MARKETING & RETAIL ANALYTICS

Project by

Jaigururam G S

AGENDA

Part A:

Customer Buying Patterns & Marketing Strategies.

Part B:

Point of Sale Data Analysis & Revenue Enhancement

PART A

1. Problem Statement
2. Data Overview
    - Dataset
    - Data Dictionary
    - Data
3. Data Assumptions
    - Data Quality & Completeness
    - Business Operations
    - Sales Pattern
    - Product Details
    - Order Processing
    - Customer
4. Exploratory Data Analysis
    - Univariate Analysis
    - Bi - Variate Analysis
    - Multi – Variate Analysis
    - Time – Series Analysis
5. Customer Segmentation Using RFM
    - What is RFM?
    - Parameters & Assumptions
    - KNIME Workflow Image
    - Output Table
6. Interferences from RFM & Identified Segments
    - Best Customers
    - Customers on Verge of Churning
    - Lots Customers
    - Loyal Customers
7. Recommendation

PROBLEM STATEMENT

An automobile parts manufacturing company has collected data on transactions for 3 years. They do not have any in-house data science team; thus, they have hired you as their consultant. Your job is to use your data science skills to find the underlying buying patterns of the customers, provide the company with suitable insights about their customers, and recommend customized marketing strategies for different segments of customers.

Dataset: Sales_Data.xlsx

DATA OVERVIEW

1. **Dataset**

   Dataset Contains two sheets namely ['Data', 'Data Dictionary ']
   - Data: Likely contains the transactional data related to the automobile parts company's sales.
   - Data Dictionary: Provides metadata about the dataset's columns.
2. **Data**
   - Dataset was passed through the “Excel Reader Node” and it contains 2,747 rows & 20 columns. There are no missing values in the dataset.
   - Insights from the “Data Explorer Node” are below:
     - **Data Type Data:**
       - ORDERDATE is Data Type column.
     - **Categorical Data:**
       - There are 12 categorical columns.
       - STATUS columns had 6 unique values:
         - Shipped,
         - Cancelled,
         - Resolved,
         - On Hold,
         - In Process,
         - Disputed

DATA OVERVIEW

- PRODUCT LINE column had 7 unique values:
  - Classic Cars,
  - Vintage Cars,
  - Motorcycles,
  - Planes,
  - Trucks and Buses,
  - Ships,
  - Trains
- DEALSIZE column had 3 unique values:
  - Medium,
  - Small,
  - Large
- PRODUCTCODE, CUSTOMERNAME, PHONE, ADDRESSLINE1 , CITY, POSTALCODE, COUNTRY, CONTACTLASTNAME, CONTACTFIRSTNAME had more than 19 unique values.

- **Numerical Data:** There are 7 Numerical Variables and below are the statistics summary

Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
ORDERNUMBER	10100	10425	10259.76	91.88	8441.48	-0.01
QUANTITYORDERED	6	97	35.10	9.76	95.30	0.37
PRICEEACH	26.88	252.87	101.10	42.04	1767.58	0.70
ORDERLINENUMBER	1	18	6.49	4.23	17.90	0.58
SALES	482.13	14082.80	3553.05	1838.95	3381751.45	1.16
DAYS_SINCE_LASTORDER	42	3562	1757.09	819.28	671220.66	-0.00
MSRP	33	214	100.69	40.11	1609.20	0.58

DATA OVERVIEW

3. Data Dictionary

Column Name	Description
ORDERNUMBER	This column represents the unique identification number assigned to each order.
QUANTITYORDERED	It indicates the number of items ordered in each order.
PRICEEACH	This column specifies the price of each item in the order.
ORDERLINENUMBER	It represents the line number of each item within an order.
SALES	This column denotes the total sales amount for each order, which is calculated by multiplying the quantity ordered by the price of each item.
ORDERDATE	It denotes the date on which the order was placed.
DAYS_SINCE_LASTORDER	This column represents the number of days that have passed since the last order for each customer. It can be used to analyze customer purchasing patterns.
STATUS	It indicates the status of the order, such as "Shipped," "In Process," "Cancelled," "Disputed," "On Hold," or "Resolved"
PRODUCTLINE	This column specifies the product line categories to which each item belongs.
MSRP	It stands for Manufacturer's Suggested Retail Price and represents the suggested selling price for each item.
PRODUCTCODE	This column represents the unique code assigned to each product.
CUSTOMERNAME	It denotes the name of the customer who placed the order.
PHONE	This column contains the contact phone number for the customer.
ADDRESSLINE1	It represents the first line of the customer's address.
CITY	This column specifies the city where the customer is located.
POSTALCODE	It denotes the postal code or ZIP code associated with the customer's address.
COUNTRY	This column indicates the country where the customer is located.
CONTACTLASTNAME	It represents the last name of the contact person associated with the customer.
CONTACTFIRSTNAME	This column denotes the first name of the contact person associated with the customer.
DEALSIZE	It indicates the size of the deal or order, which are the categories "Small," "Medium," or "Large."

DATA ASSUMPTIONS

Data Quality & Completeness

- The dataset appears clean with no missing values across 2,747 rows and 20 columns.
- All order numbers are unique (ranging from 10100 to 10425), suggesting no duplicate transactions.
- The data spans 3 years based on problem statement, providing sufficient historical information for pattern analysis.

Business Operations-

- The company operates across 19 countries globally (COUNTRY field).
- They have a structured product categorization and had 7 distinct product lines (PRODUCTLINE filed).
- They maintain standardized pricing (MSRP field).

Sales Patterns

- Order sizes vary significantly from 6 to 97 units (QUANTITYORDERED field).
- Deal sizes are categorized into three segments Small, Medium, Large.
- Sales values show high variance, ranges from $482.13 to $14,082.80 with positive skewness (1.16), indicating some high-value outlier transactions.
- Customer repeat purchase patterns vary widely from 42 to 3,562 days (DAYS_SINCE_LASTORDER field).

Product Details

- Focus on automobile parts across 7 main categories.
- Price points vary considerably from $26.88 to $252.87 (PRICEEACH field).
- Classic Cars and Vintage Cars appear to be major product lines based on the categories.

Order Processing

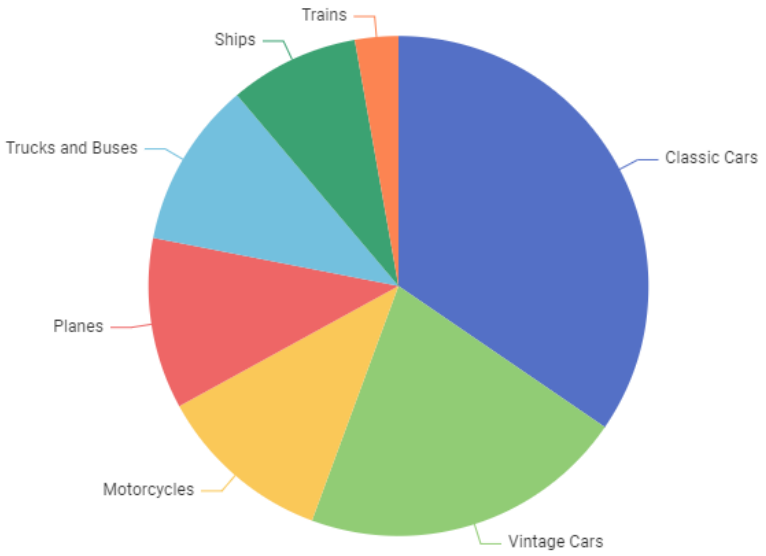
- Company has a structured order management system and 6 different STATUS categories.
- Orders can have multiple items (ORDERLINENUMBER field).
- Most orders appear to be successfully processed (STATUS: Shipped), with some exceptions.

Customer

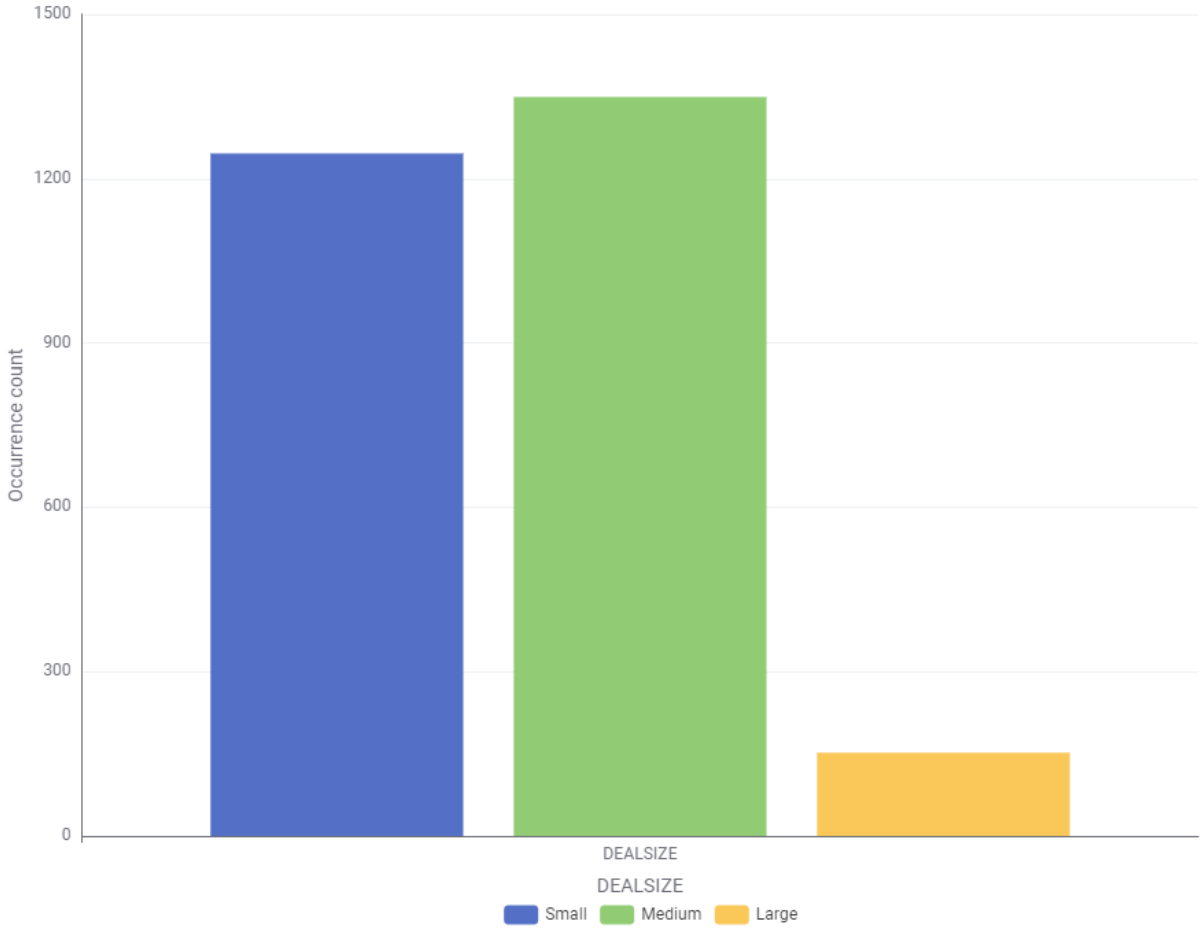
- International customer presence (COUNTRY field).
- Formal business relationships (presence of contact person details).

EDA & INFERENCES – UNIVARIATE ANALYSIS

Product line concentration



Deal size Proportions



From the pie chart for the product line, we can conclude that

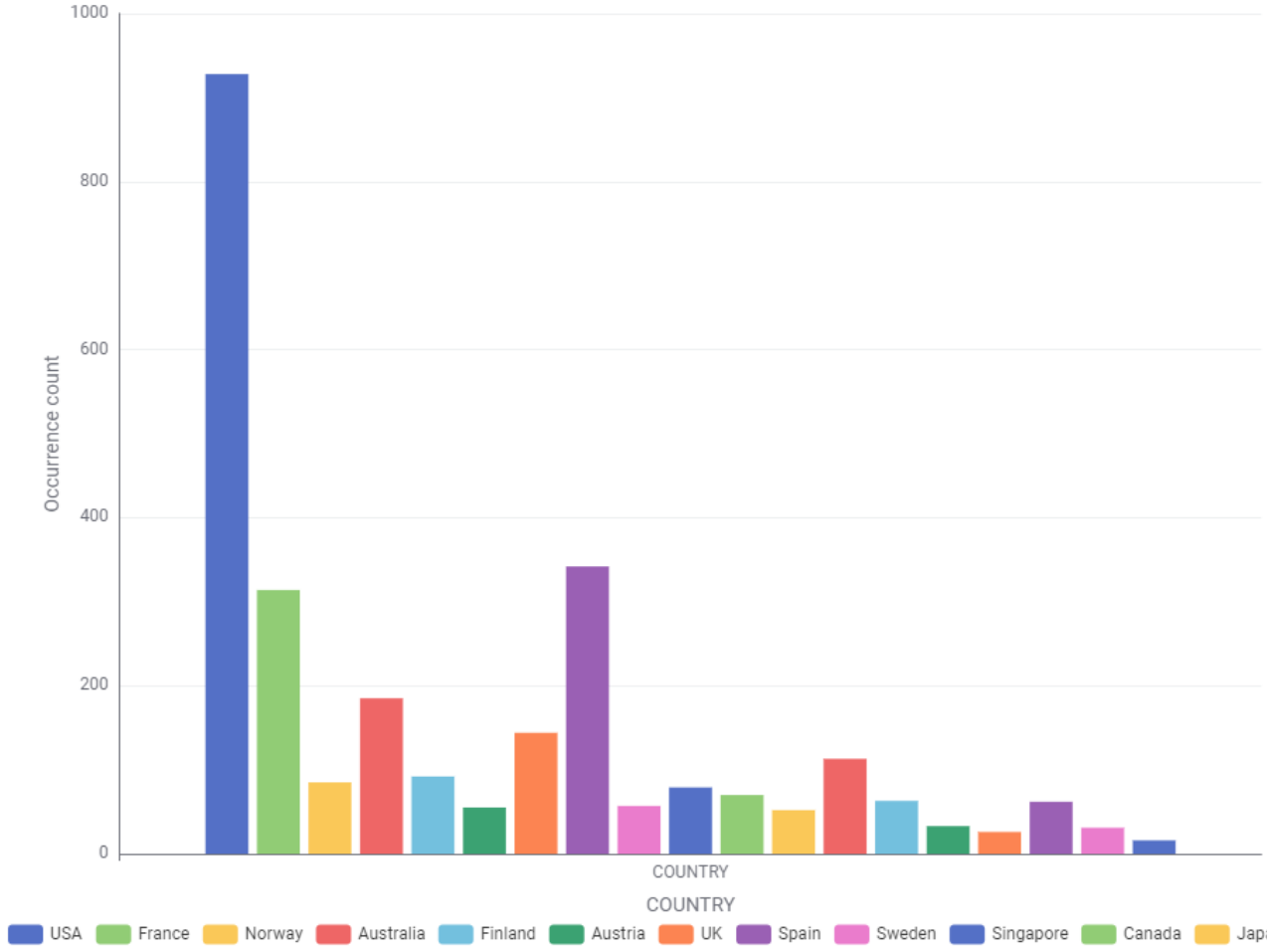
- Classic cars are the popular product followed by vintage cars and other.
- As more of customer concentration is for classic cars so we must focus on that for business growth

From the bar chart of Deal Size, we can conclude that

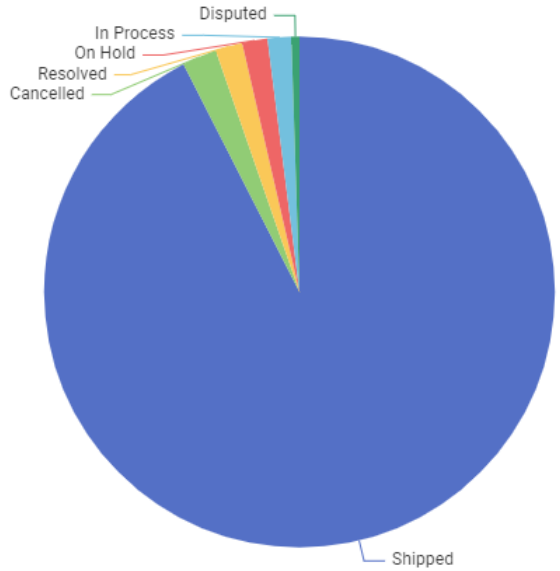
- More business is coming from the medium and small size deals.
- Its seems that customer choose the large sized deals are very less comparing to the small and medium sized deals.

EDA & INFERENCES – UNIVARIATE ANALYSIS

Country based Customer Distribution



Order Status Analysis



From the Bar chart for customer distribution globally, we can conclude that

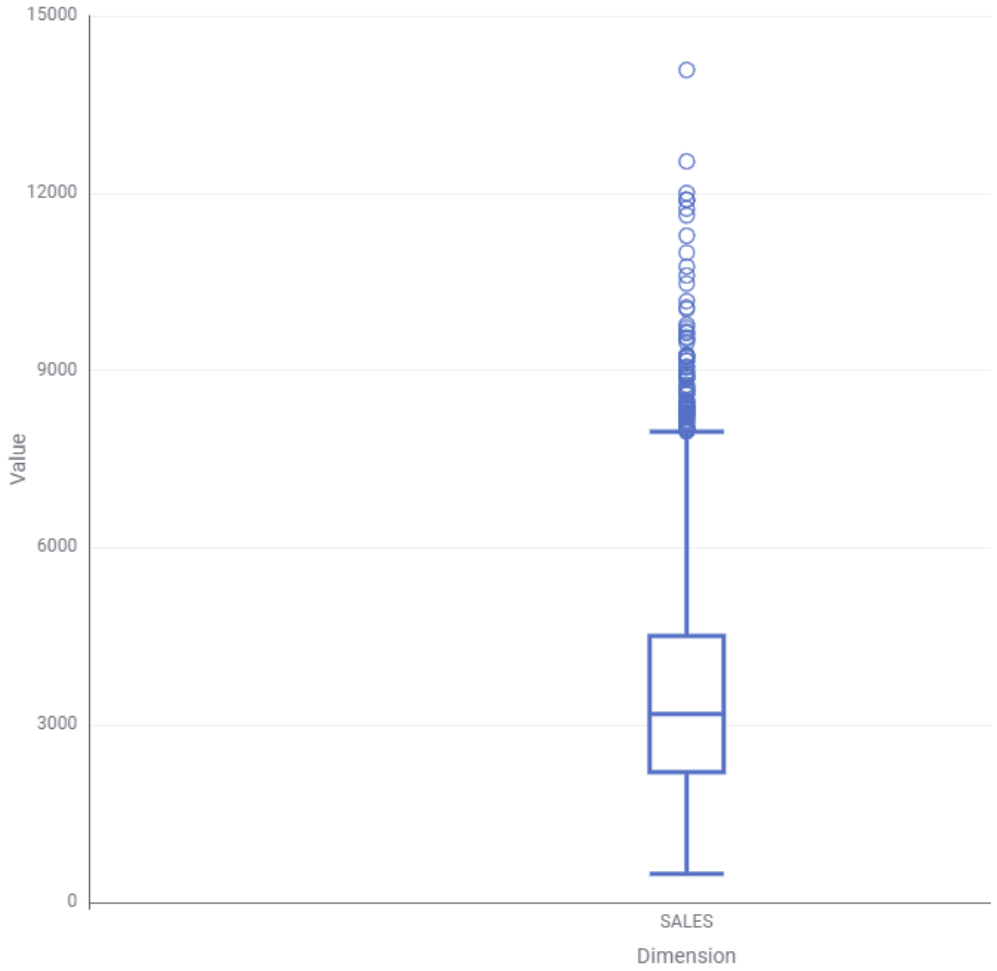
- Large number of customers are from USA followed by Spain, France and other countries.
- Doing Business across 19 countries.

From the Pie chart of order status, we can conclude that

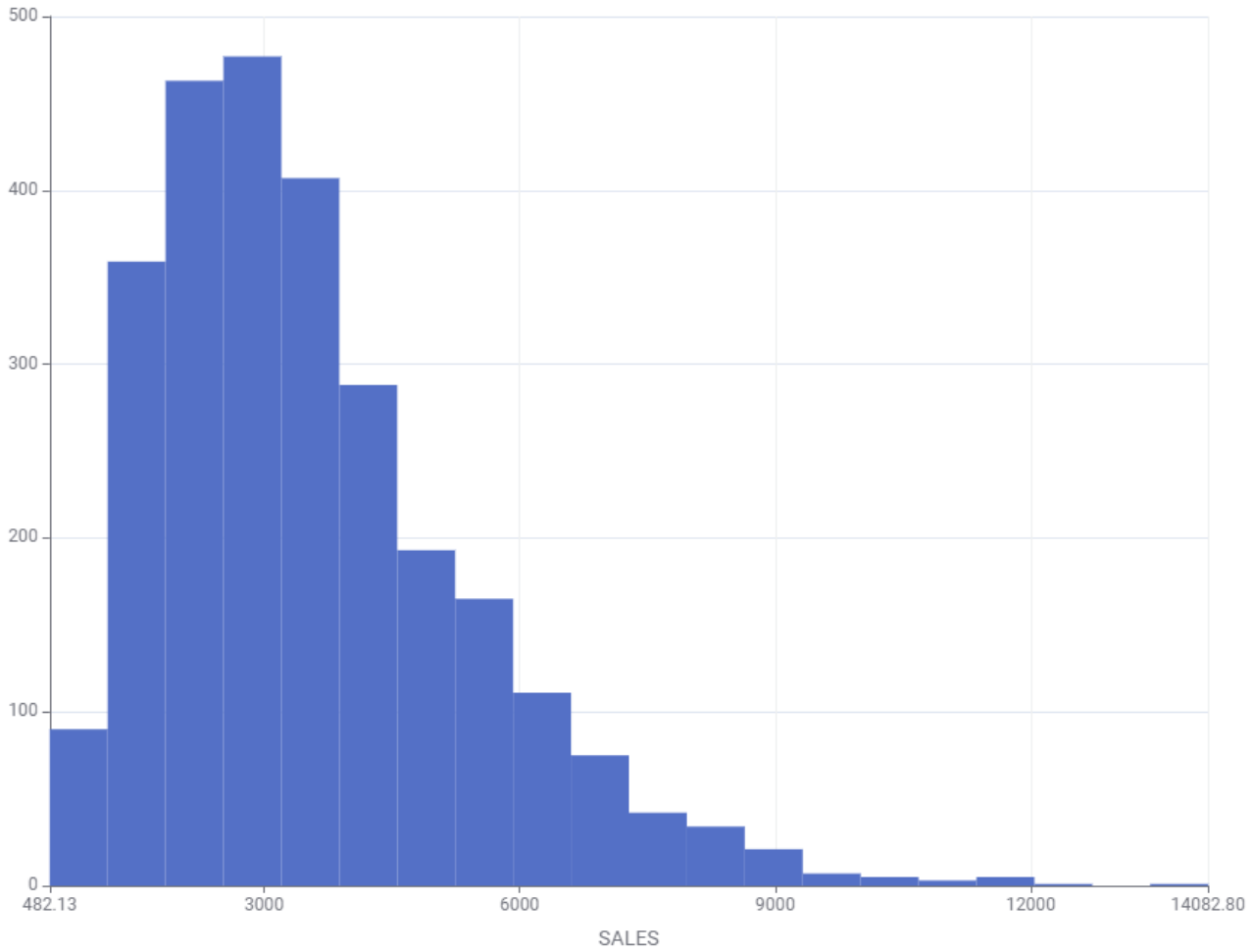
- Most of order are shipped but still cancelled order cover 2.19% of total order.
- We must focus on resolved and hold order to speed the process so that it will ship to customer on time.

EDA & INFERENCES – UNIVARIATE ANALYSIS

Box Plot (Sales)



Sales Distribution



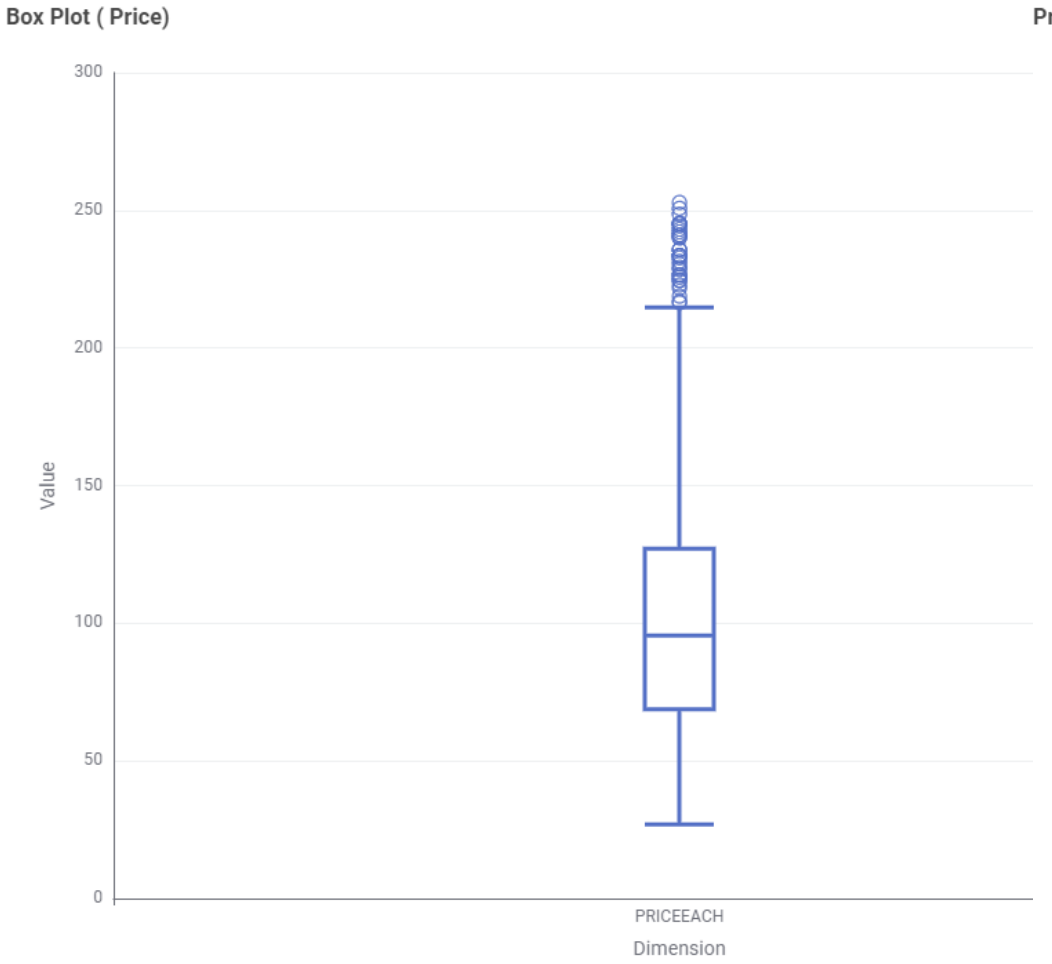
From the Box Plot for sales, we can conclude that

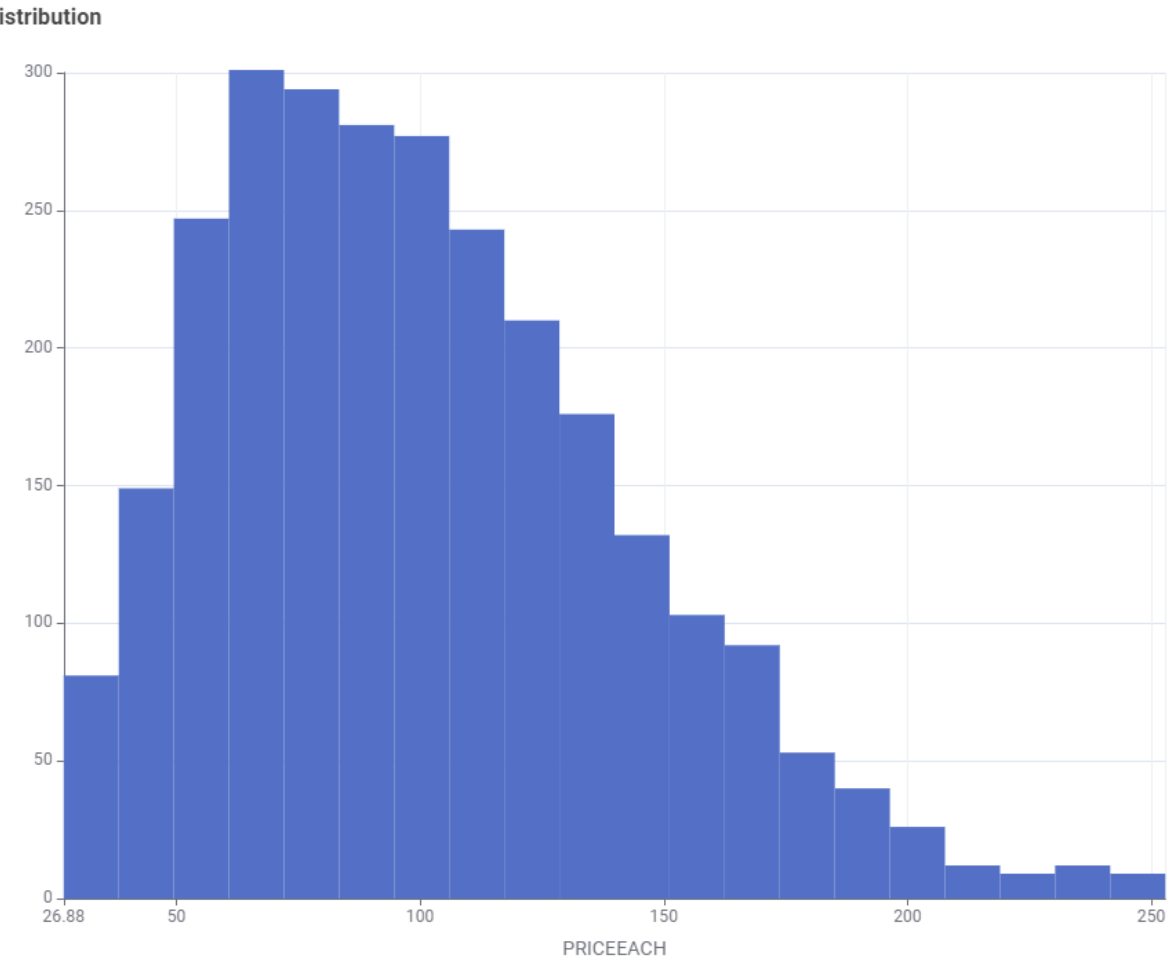
- 50% of sales are concentrated between $2.2K to $4.5K.
- There seems to be few outliers present.
- Minimum sales value was $482 and Maximum sales values was $8K.

From the Histogram sales distribution, we can conclude that

- Most of sales are concentrated from range $1.5K to $6K.
- There seems to be few outliers present

EDA & INFERENCES – UNIVARIATE ANALYSIS





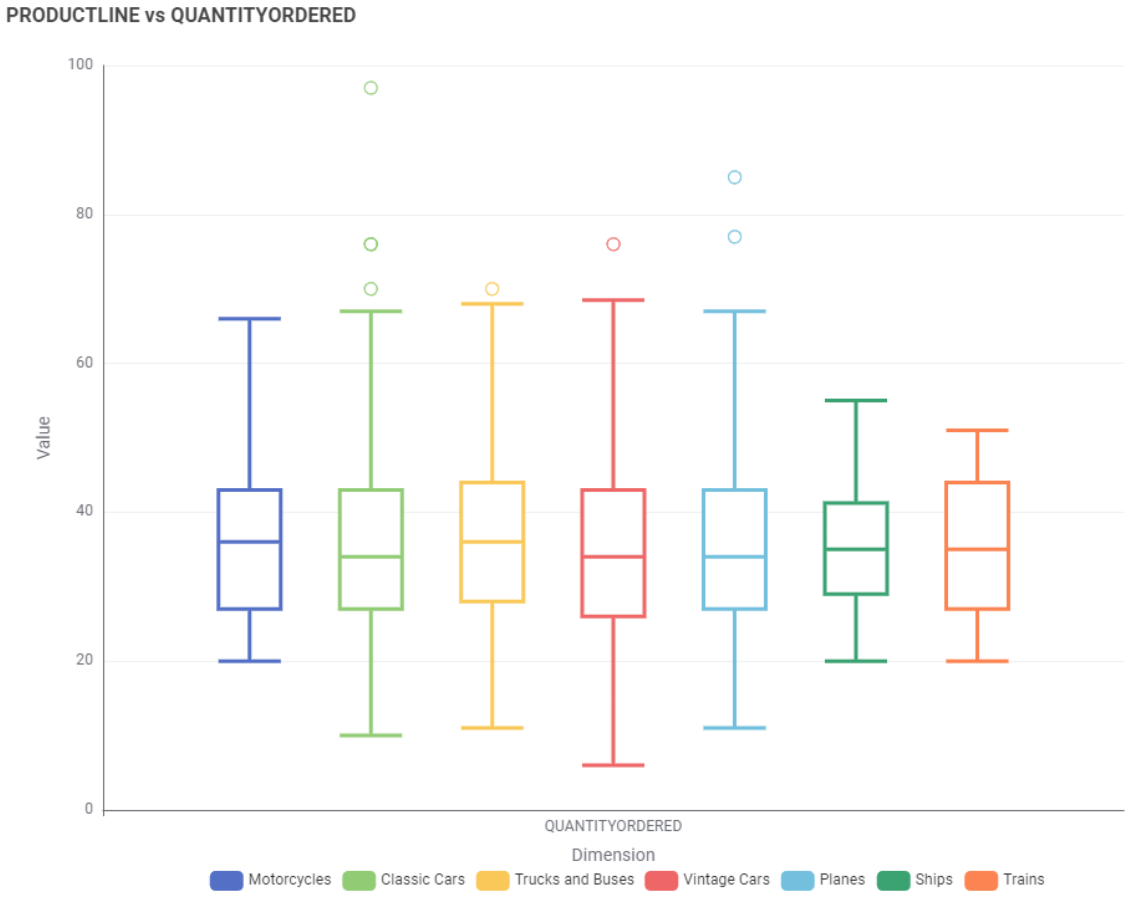
From the Box Plot for Price value, we can conclude that

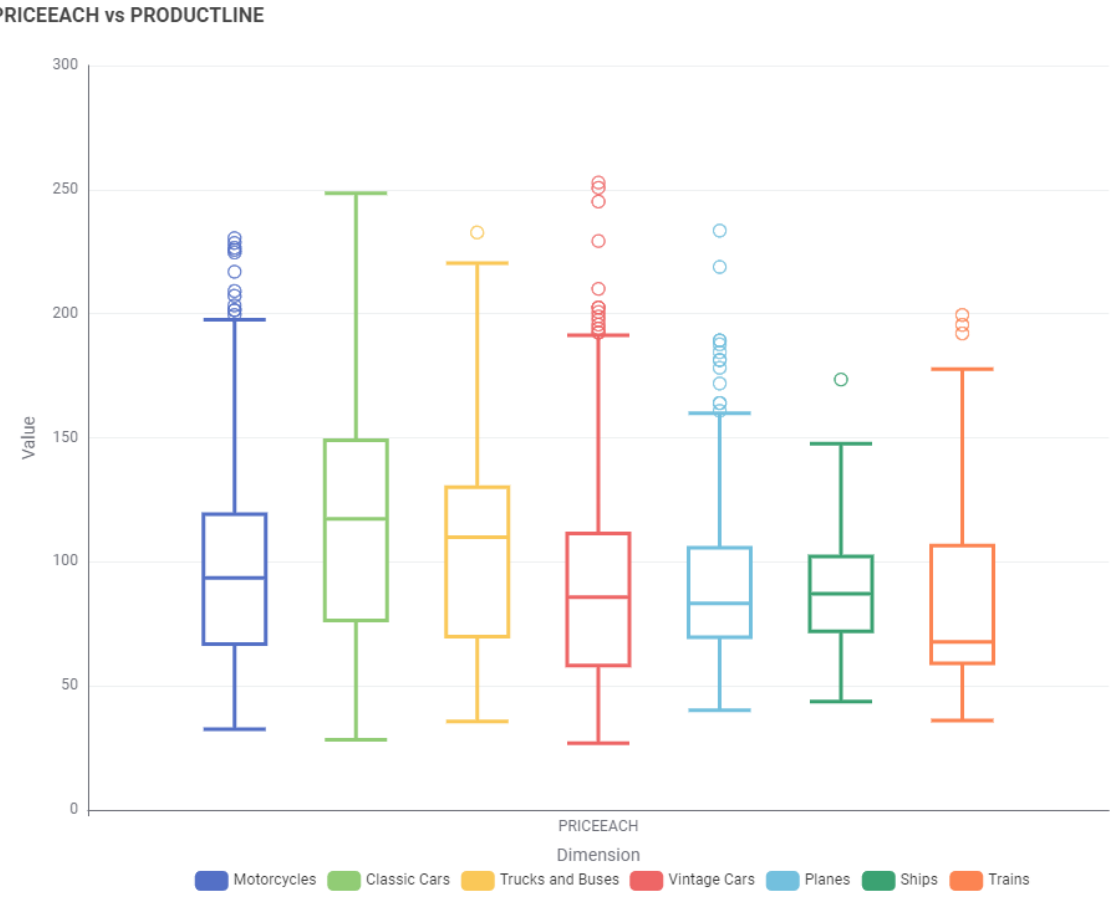
- 50% of product has price range concentrated between $68 to $127.
- There seems to be few outliers present.
- Minimum sales value was $27 and Maximum sales values was $215.

From the Histogram price distribution, we can conclude that

- Most products fall in Price range of $50 to $150.
- There seems to be few outliers present

EDA & INFERENCES – BIVARIATE ANALYSIS





From the Box Plot between Product line & quantity Ordered, we can conclude that

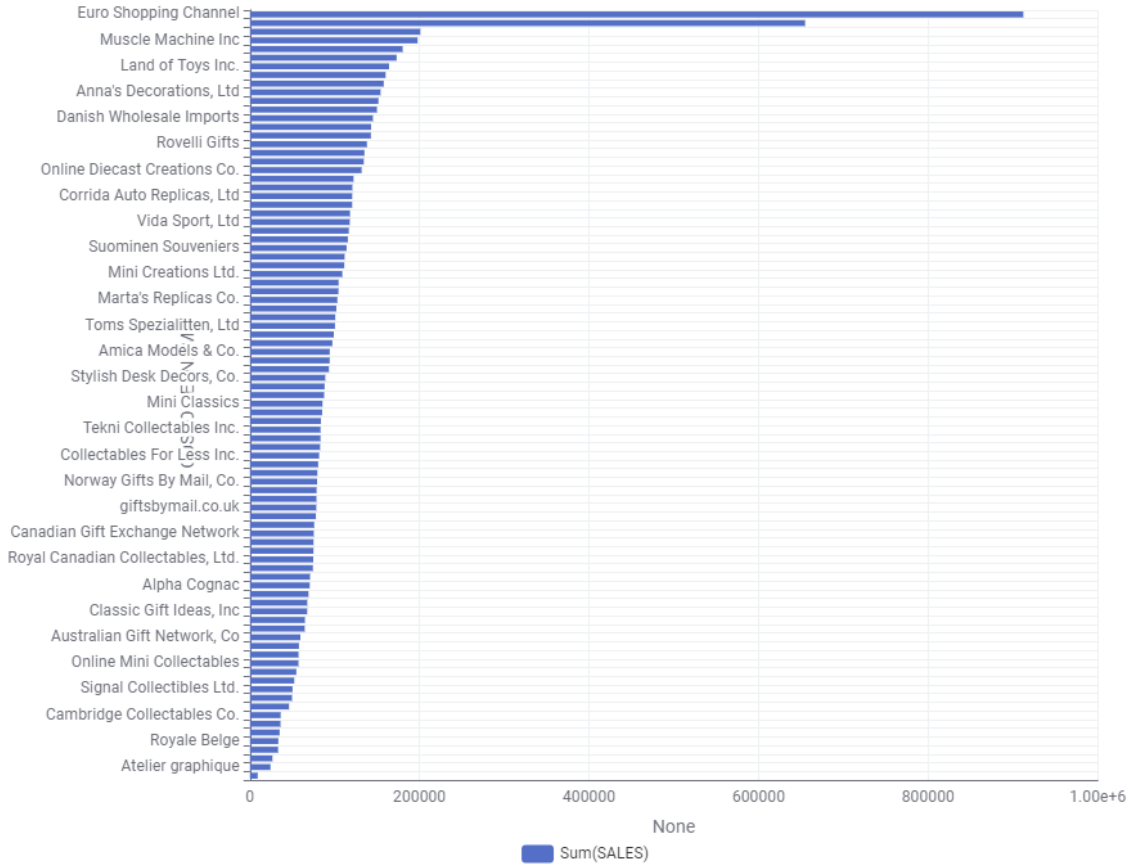
- Most product categories show similar median order quantities (around 34-36 units), this suggests relatively consistent ordering patterns across product lines
- There are notable outliers but classic cars show more significant outliers followed by planes.
- Vintage Cars and Planes show the widest spread in order quantities.

From the Box Plot between Product line & Price each, we can conclude that

- Product categories shows different median prices, this suggests non consistent pricing across product line.
- Vintage Cars, motorcycles and Planes show the widest spread in price value.
- There are notable outliers in a product category except classic cars.

EDA & INFERENCES – BIVARIATE ANALYSIS

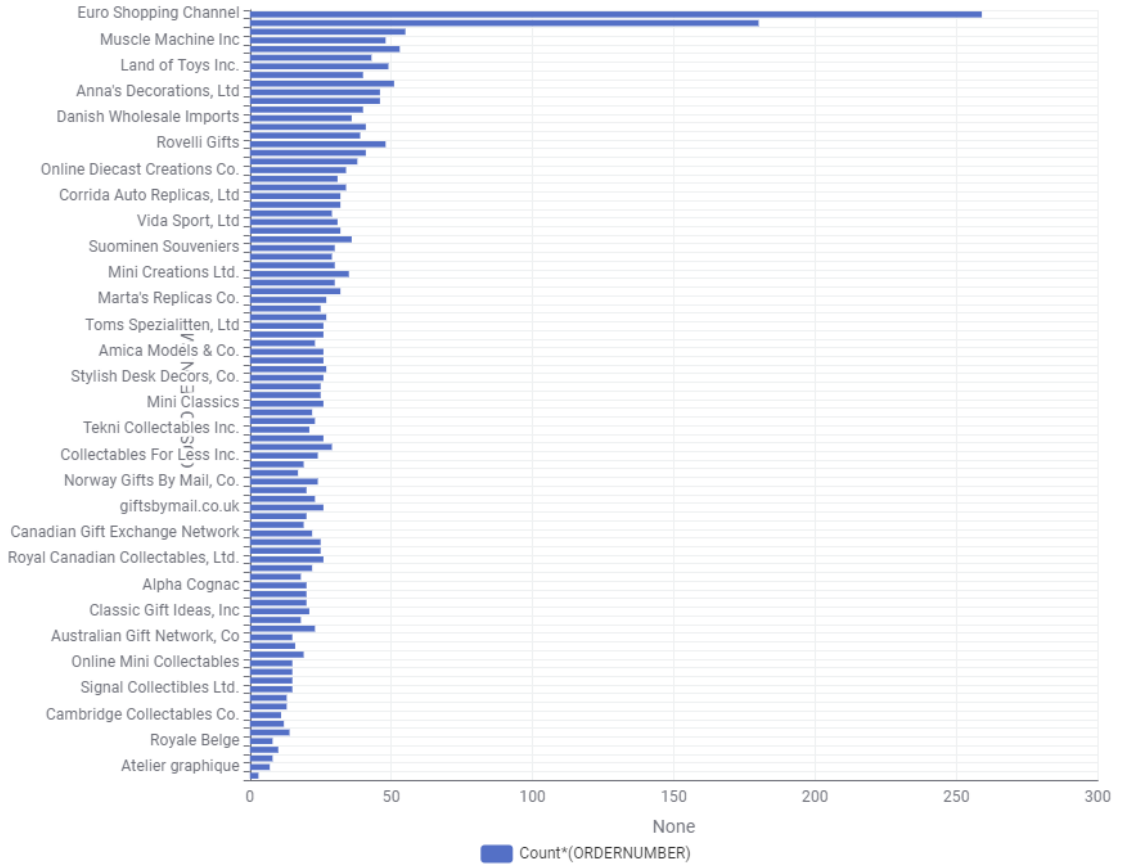
CUSTOMERNAME vs SALES



From the bar chart between customer name and sales, we can inference that

- Euro Shopping Channel was the top customer with the highest sales value followed by Mini Gifts Distributor ltd.
- Most customers had made purchase within $200K.
- Boards & Toys co. is customer which give lowest sales value.

Customer Order Frequency Analysis

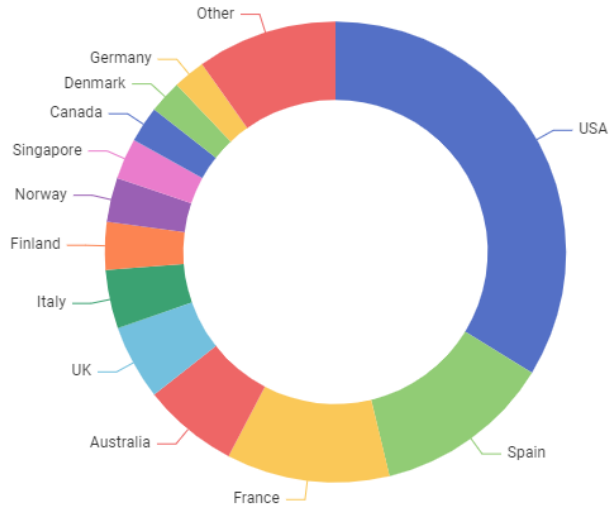


From the bar chart between Customer name and order frequency, we can inference that

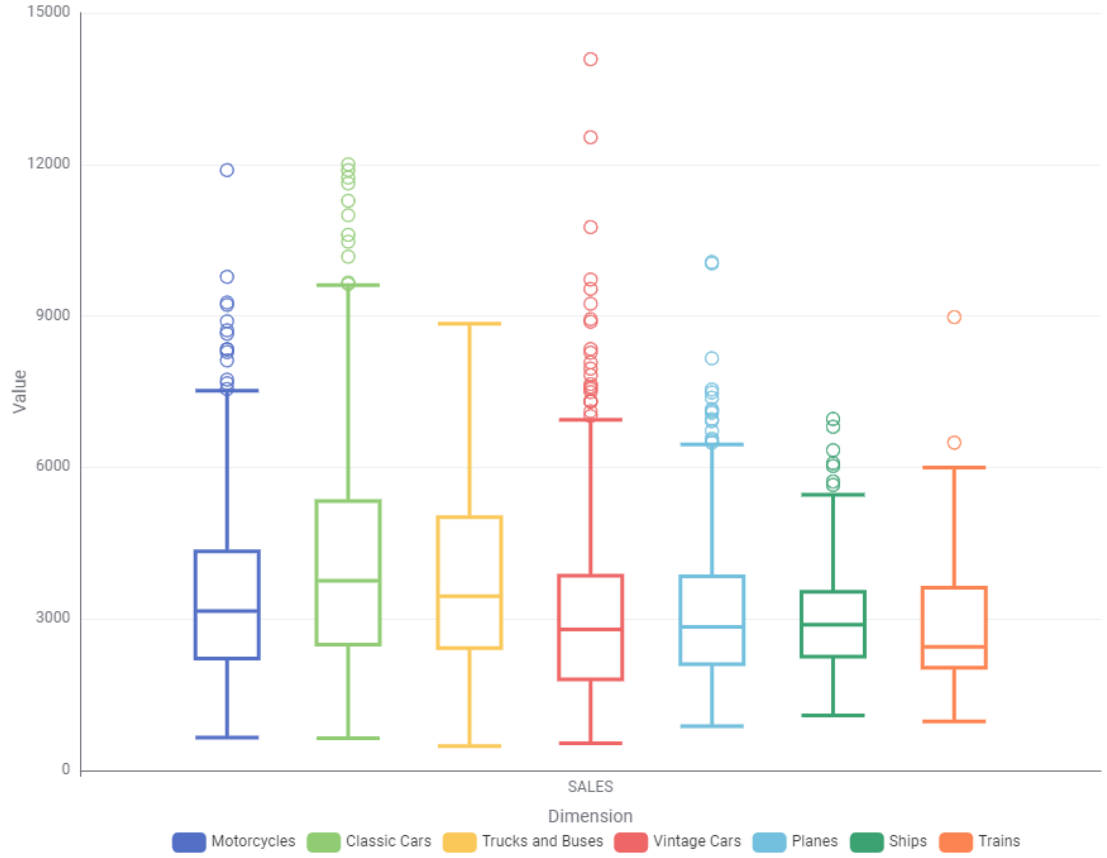
- Euro Shopping Channel was the top customer with the highest orders followed by Mini Gifts Distributor ltd.
- Order frequency range in which most customers fall under 55.
- Boards & Toys co. is customer with lowest order frequency.

EDA & INFERENCES – BIVARIATE ANALYSIS

Customer Geographic Distribution



SALES vs PRODUCTLINE



From the Pie chart of customer distribution in Country, we can conclude that

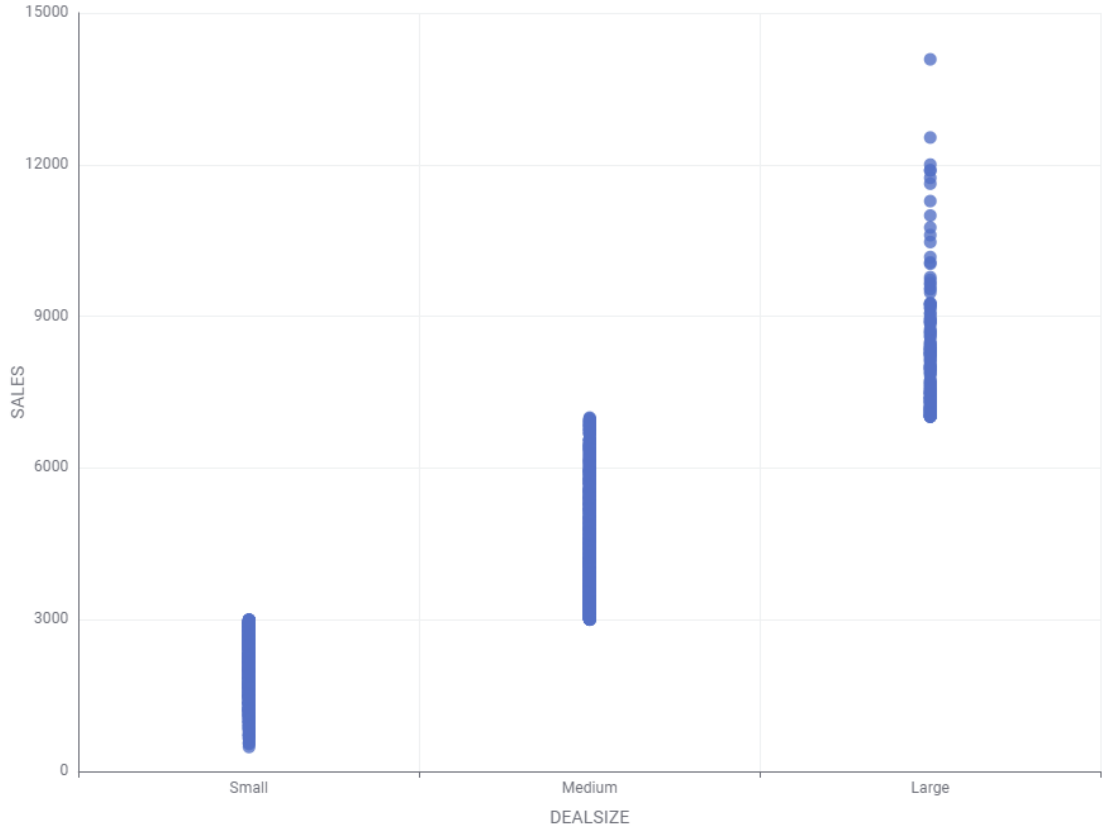
- Most of Customers belongs to USA with 34% contribution to total data.
- There are seven countries that contribute Less than 2%.
- Spain Contribute 12.5% and France Contribute 11.4%.
- USA, Spain & France cover 50% of customer base.

From Box plot between Sales & Product line, we can conclude that

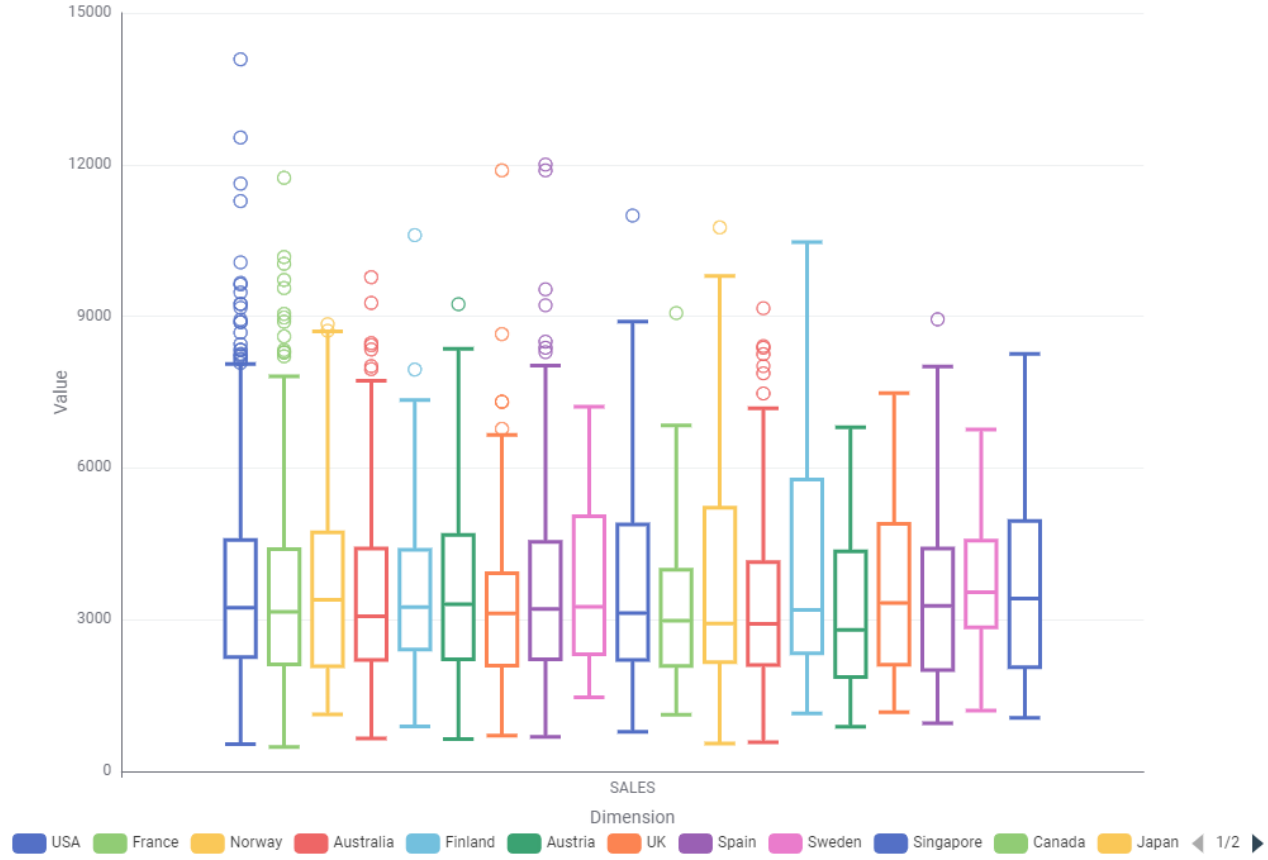
- Classic Cars shows the highest median sales and most frequent high-value transactions.
- Focus marketing efforts on Classic Cars and Vintage Cars for maximum revenue impact.
- The high outliers in Vintage Cars suggest potential for premium pricing strategies.
- Planes and Ships have more concentrated sales distributions, suggesting more consistent pricing or order patterns.

EDA & INFERENCES – BIVARIATE ANALYSIS

SALES vs DEALSIZE



SALES vs COUNTRY



From Scatterplot between Sales & Deal size, we can conclude that

- There appears to be a positive correlation between deal size and sales.
- The business might want to focus on increasing the number of large deals, as they seem to generate higher sales volumes.
- There are a few high-performing large deals that stand out above the rest, which might be worth analyzing for replication.
- 3 distinct deal sizes: Small, Medium, & Large.

From Box plot between Sales & Country, we can conclude that

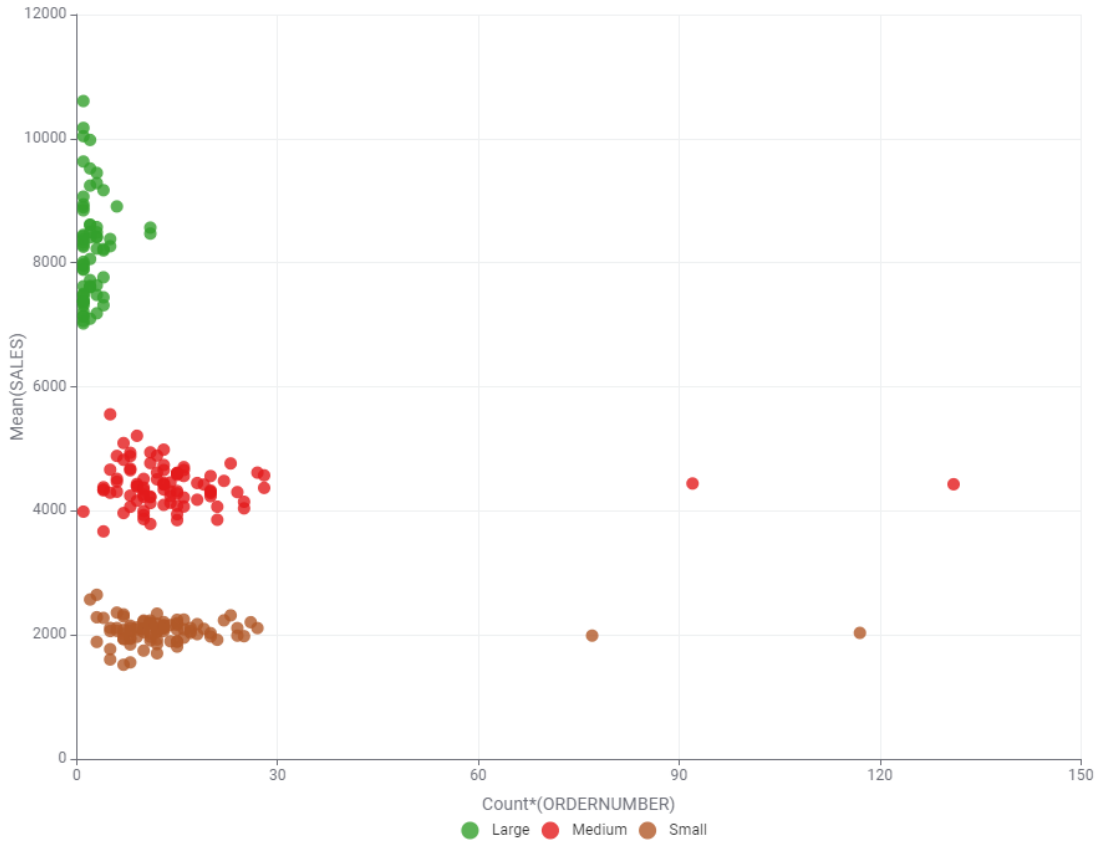
• Some countries, such as the USA, France, and Norway, show higher maximum sales values.

• Several countries have outliers, suggesting exceptional sales performances in these markets.

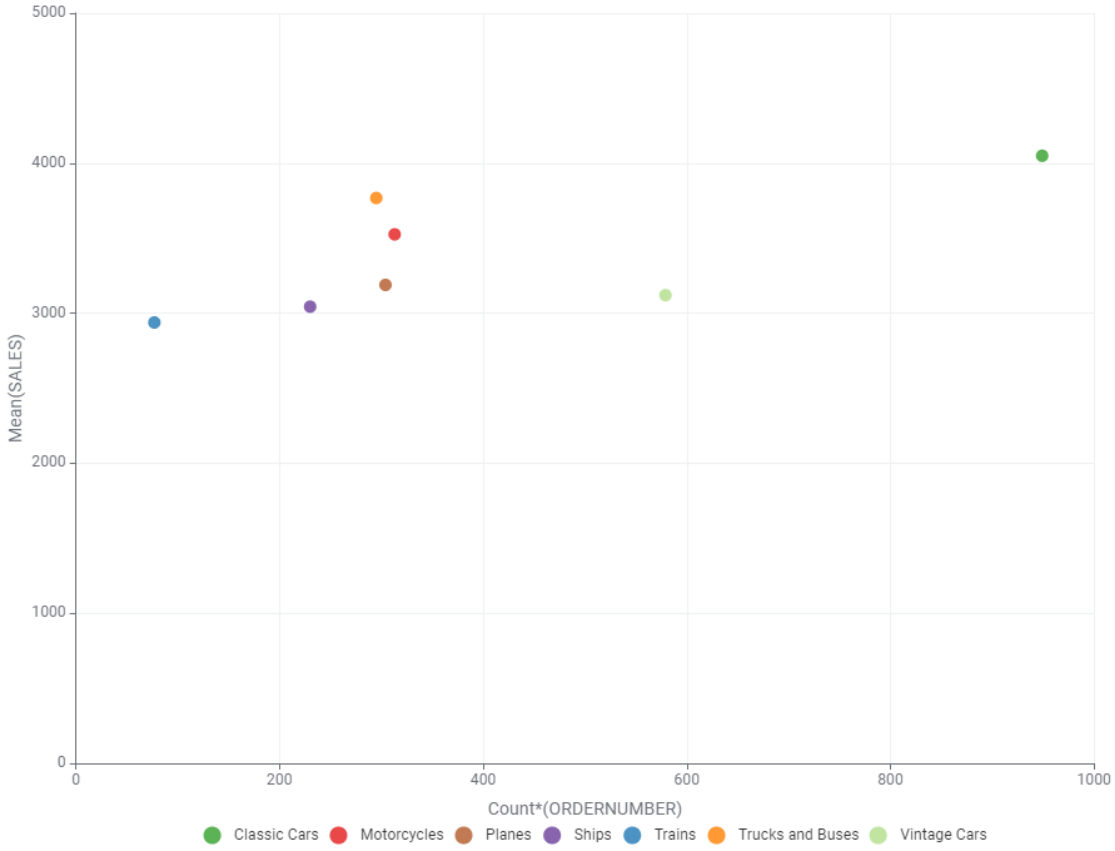
• Countries with higher max sales but lower medians (e.g., Singapore, Canada) might represent potential for growth.

EDA & INFERENCES – MULTIVARIATE ANALYSIS

Customer Order Behavior



Product Performance Deep Dive



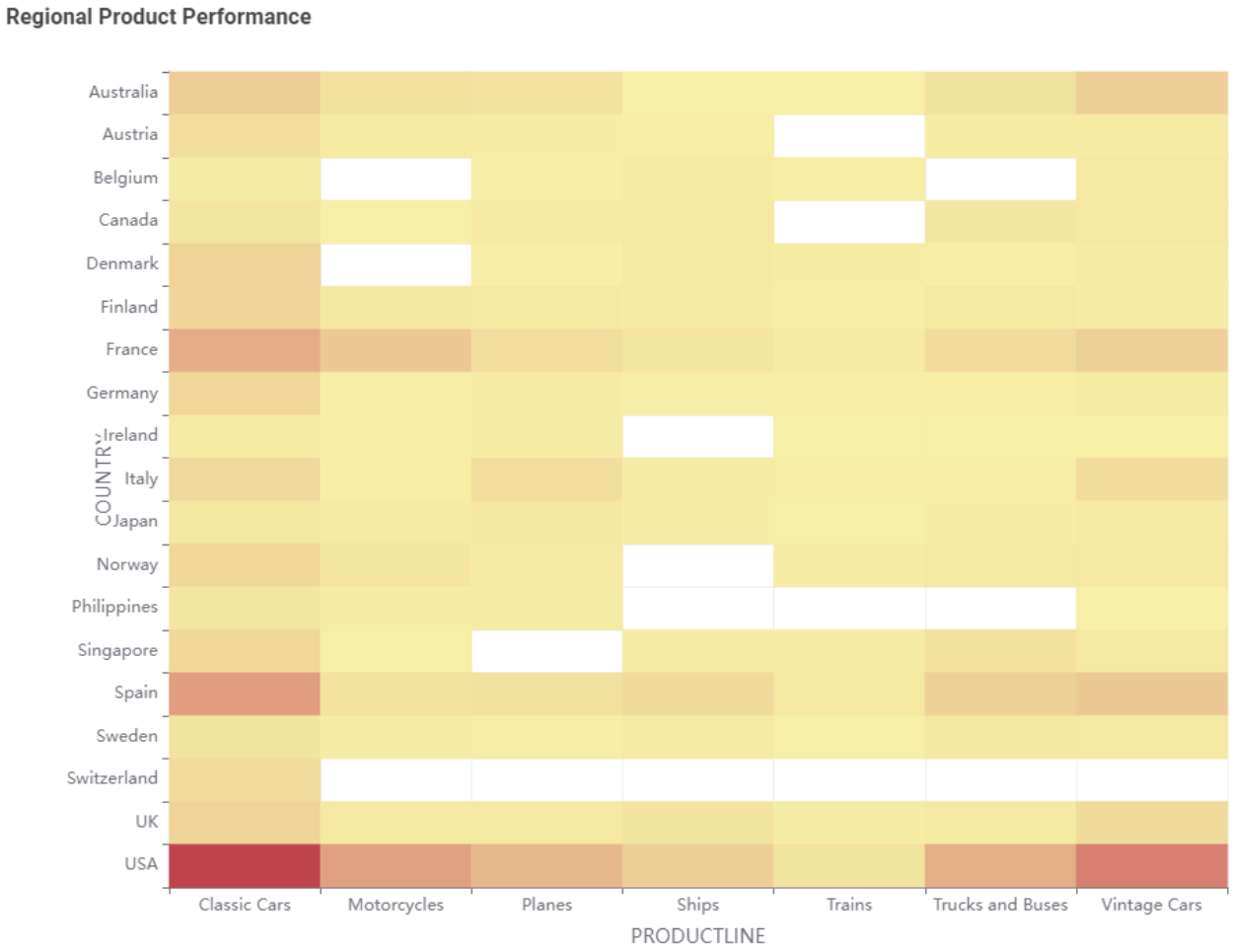
From Customer Order Behavior plot, insights are

- Large orders have the highest mean sales but occur less frequently. Small orders have the lowest mean sales but are the most numerous.
- The majority of high value orders come from customers with fewer total orders, suggesting that big spenders aren't necessarily frequent buyers.
- A diverse customer base that includes both high-value and high frequency purchasers

From Product Performance Deep-Dive plot, insights are

- Classic Cars have the highest number of orders, indicating strong customer demand.
- Motorcycles show a good balance between order volume & mean sales value, potentially representing a stable & profitable product line.
- Despite lower order volumes, categories like Vintage Cars, Trucks & Buses show high mean sales values, suggesting they may be premium items.

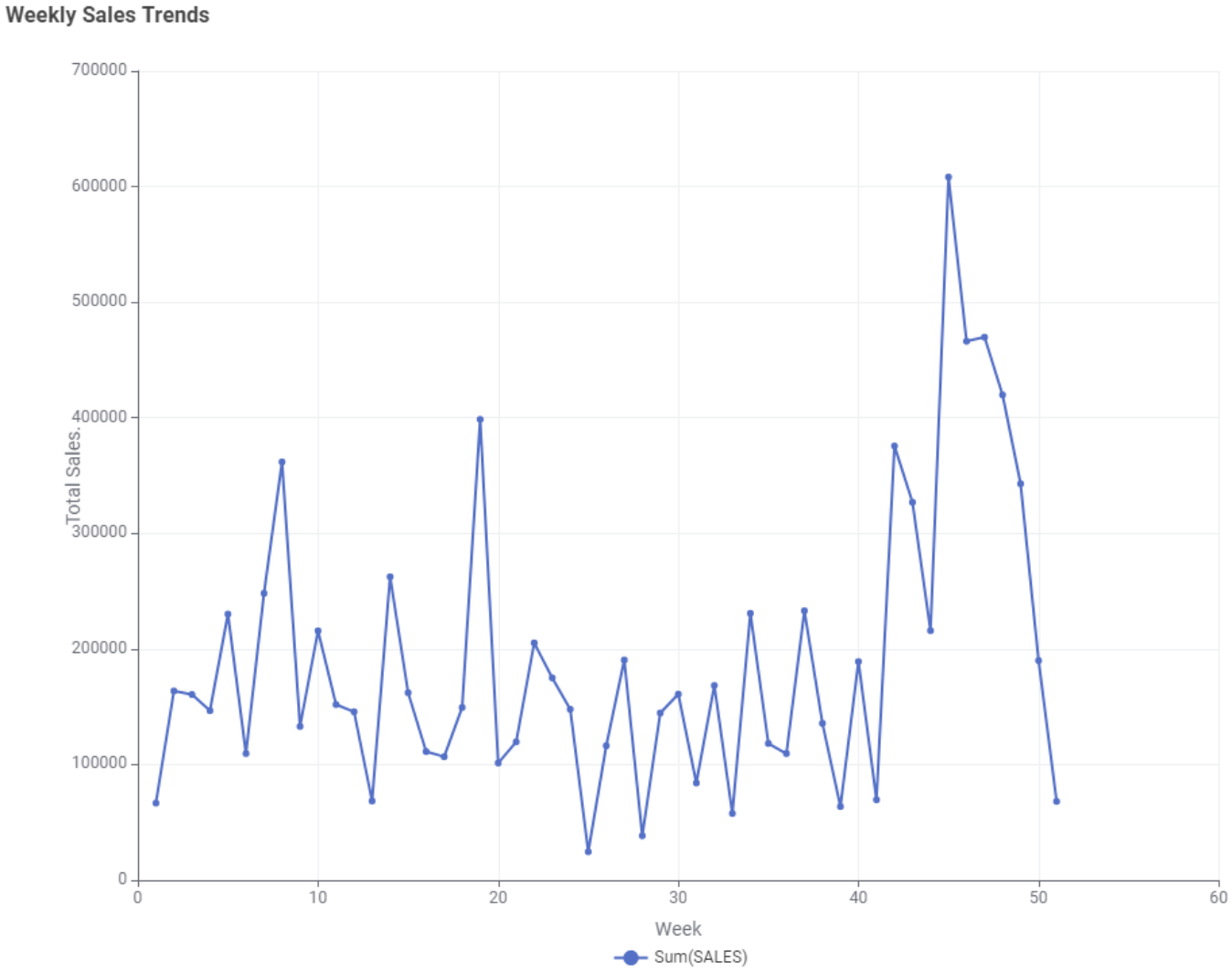
EDA & INFERENCES – MULTIVARIATE ANALYSIS



From Regional Product Performance plot, insights

- The USA stands out as the strongest market across all product lines.
- Classic Cars appear to be the best performing product category across most countries, indicating a globally popular product line.
- European countries like France, Spain, & Germany show strong performance across multiple product categories.
- Several countries show no data for certain product lines, suggesting potential markets.

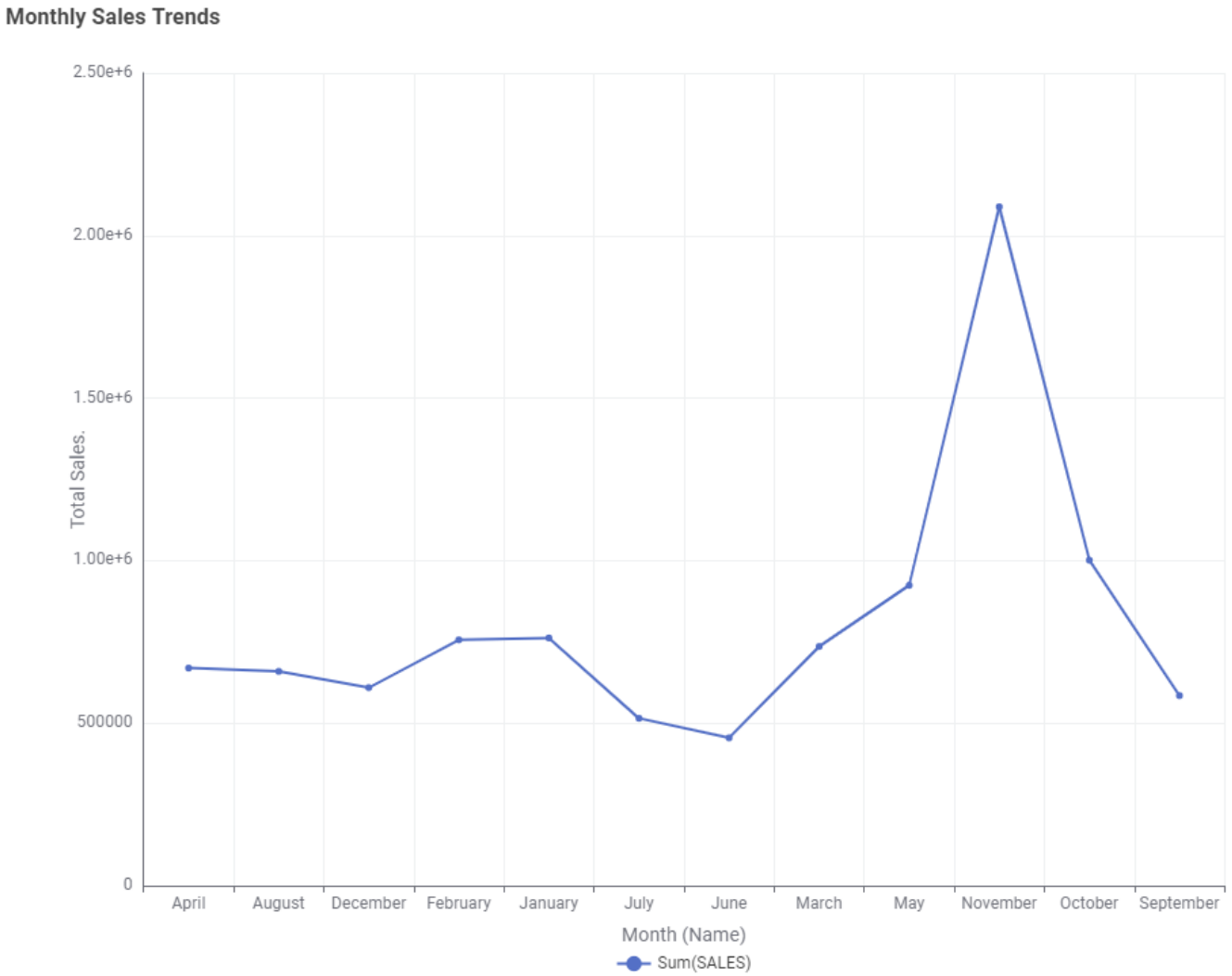
EDA & INFERENCES –TIME SERIES ANALYSIS



Insights form Weekly Sales Trends plot:

- Sales show extreme fluctuations week to week, indicating an unstable or highly seasonal business.
- The highest sales peak occurs around week 45. This suggests a potential seasonal high point or successful marketing campaign.
- There seems to be a repeating pattern of peaks and troughs, possibly aligned with monthly or quarterly cycles.

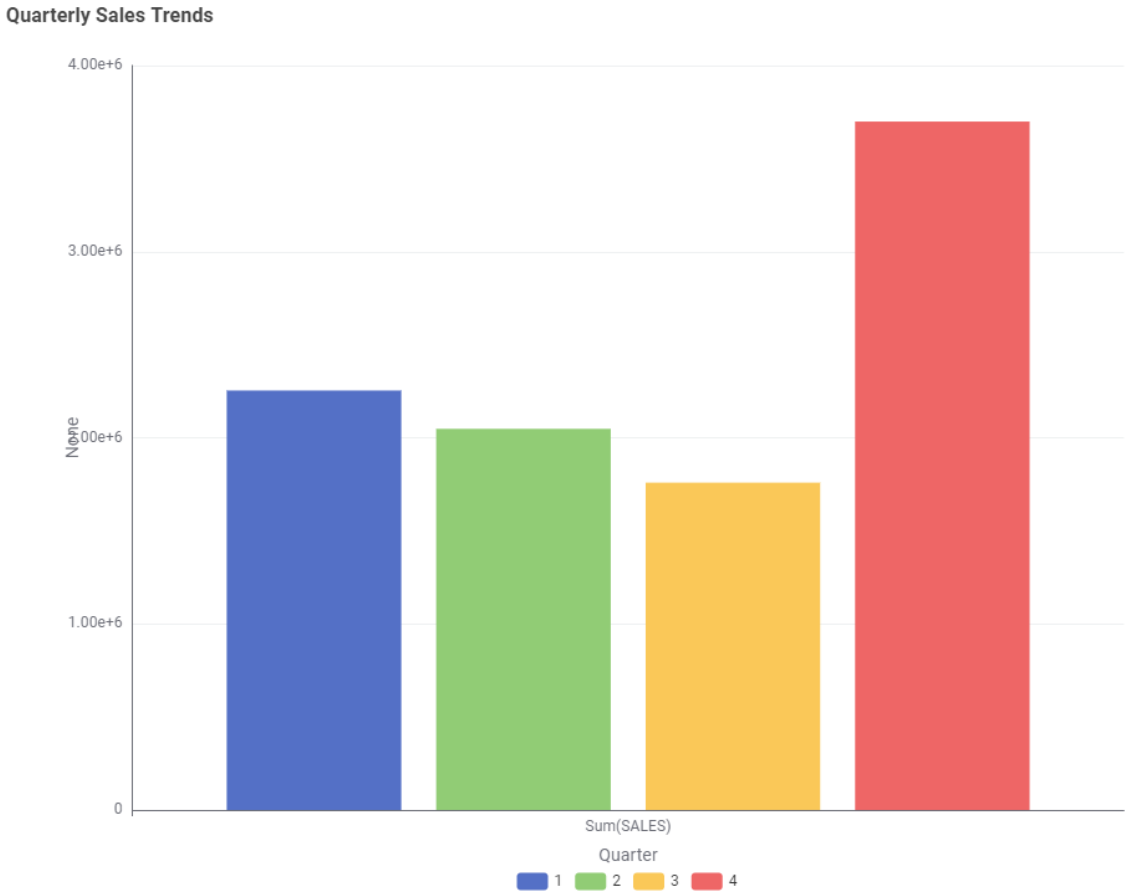
EDA & INFERENCES –TIME SERIES ANALYSIS

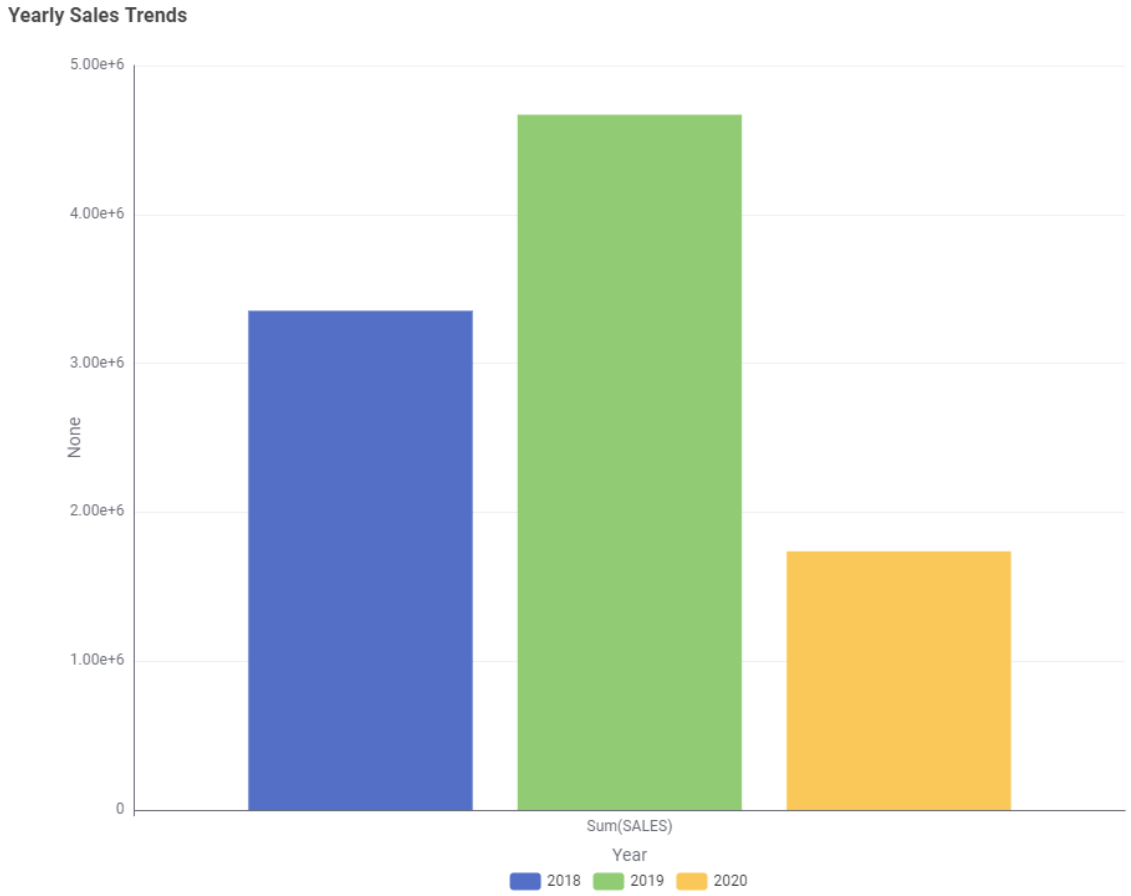


Insights form Monthly Sales Trends plot:

- There's a significant sales spike in November compared to other months. This suggests a strong seasonal trend, possibly due to holiday shopping.
- June & July show the lowest sales figures, pointing to a potential off-season that may require targeted marketing or product diversification.
- From July to November, there's a steady increase in sales.

EDA & INFERENCES –TIME SERIES ANALYSIS





Insights form Quarterly Sales Trends plot:

- The fourth quarter significantly outperforms all others, with sales nearly double that of Q1. This indicates a strong seasonal trend, likely driven by holiday shopping.
- Q3 shows the lowest sales, suggesting a potential off-season that may require targeted strategies to boost performance.

Insights form Yearly Sales Trends plot:

• Sales reached their highest point in 2019, showing significant growth from 2018.

• There's a sharp drop in sales in 2020, falling below even 2018 levels. This could be due to external factors.

EDA & INFERENCES –TIME SERIES ANALYSIS

Conclusion:

• The business shows consistent seasonal trends but is vulnerable to yearly fluctuations. Success depends on maximizing Q4 sales, smoothing off peak periods, and building resilience against economic shocks. Strategies may include diversification, market expansion, and year-round customer engagement.

CUSTOMER SEGMENTATION USING RFM ANALYSIS

What is RFM?

RFM stands for Recency, Frequency, and Monetary analysis. It is a customer segmentation technique used to evaluate and group customers based on their purchasing behavior:

- **Recency (R):** How recently a customer has made a purchase.
- **Frequency (F):** How often a customer makes purchases.
- **Monetary (M):** How much a customer spends.

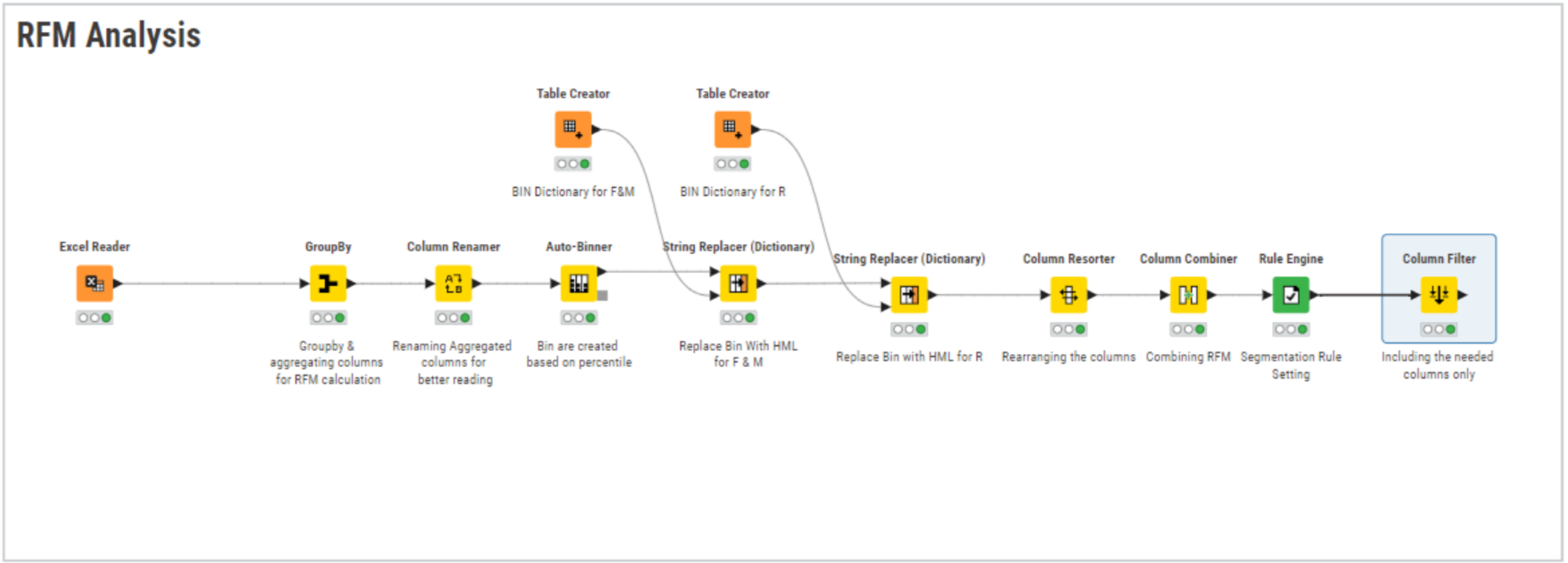
Parameters & Assumptions

Used KNIME tool for RFM Analysis and also include all customers provided in dataset.

1. **Recency:** Days since the last purchase (DAYS_SINCE_LASTORDER).
   - Assumption: Customers with lower values are more engaged.
2. **Frequency:** Number of orders per customer (ORDERNUMBER).
   - Assumption: Higher order counts indicate loyal customers.
3. **Monetary:** Total sales per customer (SALES).
   - Assumption: Higher spenders are more valuable.

CUSTOMER SEGMENTATION USING RFM ANALYSIS

KNIME workflow image



CUSTOMER SEGMENTATION USING RFM ANALYSIS

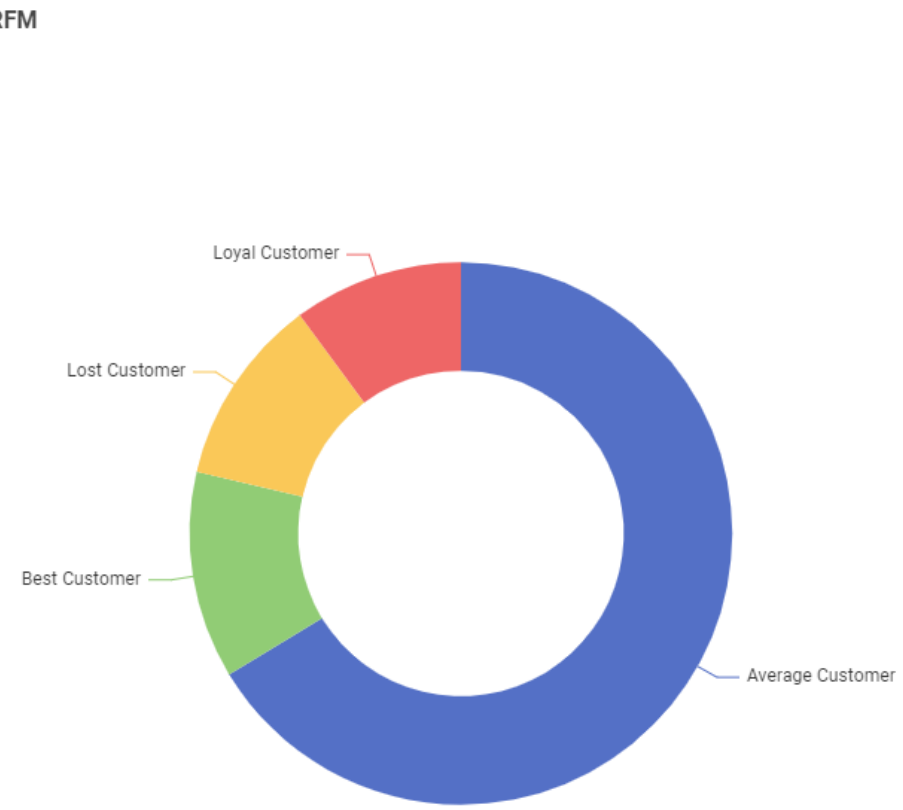
Output Table

#	RowID	CUSTOMERNA... *String*	Recency *Number (integer)*	Frequency *Number (integer)*	Monetary *Number (double)*	RFM_Segment *String*	Segmentation *String*
1	Row0	AV Stores, Co.	421	51	157,807.81	MHH	Loyal Customer
2	Row1	Alpha Cognac	675	20	70,488.44	LLL	Lost Customer
3	Row2	Amica Models & Co.	328	26	94,117.26	MMM	Average Customer
4	Row3	Anna's Decorations,...	131	46	153,996.13	HHH	Best Customer
5	Row4	Atelier graphique	312	7	24,179.96	MLL	Average Customer
6	Row5	Australian Collecta...	1018	23	64,591.46	LML	Average Customer
7	Row6	Australian Collector...	229	55	200,995.41	HHH	Best Customer
8	Row7	Australian Gift Netw...	190	15	59,469.12	HLL	Average Customer
9	Row8	Auto Assoc. & Cie.	275	18	64,834.32	MLL	Average Customer
10	Row9	Auto Canal Petit	127	27	93,170.66	HMM	Average Customer

CUSTOMER SEGMENTATION USING RFM ANALYSIS

Inferences from RFM Analysis and identified segments

After conducting the RFM analysis, we have classified customers into four segments based on their purchasing behavior.



Customer are segmented into four groups

- Best Customer
- Loyal Customer
- Average Customer
- Lost Customer

Best Customer: HHH (High Recency, Frequency & Monetary).
Loyal Customer: MHH / LHH (High Frequency & Monetary).
Lost Customer: LLL (Low Recency, Frequency & Monetary).
Average Customer: All other combination except the above mentioned.

CUSTOMER SEGMENTATION USING RFM ANALYSIS

Best Customers

These customers have high Recency, Frequency, and Monetary scores, indicating that they purchase frequently, spend generously, and have made recent transactions.

#	RowID	CUSTOMERNAME *String*	Recency *Number (integer)*	Frequency *Number (integer)*	Monetary *Number (double)*	RFM_Segment *String*	Segmentation *String*
	Ro'	CUSTOMERNAME	Recency	Frequency	Monetary	RFM_Segment	1 selected
4	Row3	Anna's Decorations, Ltd	131	46	153,996.13	HHH	Best Customer
7	Row6	Australian Collectors, Co.	229	55	200,995.41	HHH	Best Customer
33	Row...	Euro Shopping Channel	42	259	912,294.11	HHH	Best Customer
44	Row...	La Rochelle Gifts	139	53	180,124.9	HHH	Best Customer
45	Row...	Land of Toys Inc.	216	49	164,069.44	HHH	Best Customer
54	Row...	Mini Gifts Distributors Ltd.	219	180	654,858.06	HHH	Best Customer
58	Row...	Online Diecast Creations Co.	253	34	131,685.3	HHH	Best Customer
68	Row...	Salzburg Collectables	188	40	149,798.63	HHH	Best Customer
73	Row...	Souveniers And Things Co.	186	46	151,570.98	HHH	Best Customer
77	Row...	Technics Stores Inc.	241	34	120,783.07	HHH	Best Customer
79	Row...	The Sharp Gifts Warehouse	182	40	160,010.27	HHH	Best Customer

CUSTOMER SEGMENTATION USING RFM ANALYSIS

Customers on the Verge of Churnings

These customers have moderate Frequency and Monetary scores but low Recency scores, indicating that they have not made a purchase recently and might churn soon.

#	RowID	CUSTOMERNAME *String*	Recency *Number (integer)*	Frequency *Number (integer)*	Monetary *Number (double)*	RFM_Segment *String*	Segmentation *String*
	Ro'	CUSTOMERNAME	Recency	Frequency	Monetary	3 selected	Segmentation
6	Row5	Australian Collectables, Ltd	1018	23	64,591.46	LML	Average Customer
14	Row...	Blauer See Auto, Co.	705	22	85,171.59	LMM	Average Customer
25	Row...	Cruz & Sons Co.	971	26	94,015.73	LMM	Average Customer
29	Row...	Diecast Collectables	672	18	70,859.78	LLM	Average Customer
32	Row...	Enaco Distributors	659	23	78,411.86	LMM	Average Customer
47	Row...	Marseille Mini Autos	757	25	74,936.14	LMM	Average Customer
57	Row...	Norway Gifts By Mail, Co.	825	24	79,224.23	LMM	Average Customer
72	Row...	Signal Gift Stores	657	29	82,751.08	LMM	Average Customer
74	Row...	Stylish Desk Decors, Co.	702	26	88,804.5	LMM	Average Customer
83	Row...	Toys4GrownUps.com	649	30	104,561.96	LMM	Average Customer

CUSTOMER SEGMENTATION USING RFM ANALYSIS

Lost Customers

These customers have low scores for Recency, Frequency, and Monetary, indicating minimal engagement and likely Lost.

#	RowID	CUSTOMERNAME *String*	Recency *Number (integer)*	Frequency *Number (integer)*	Monetary *Number (double)*	RFM_Segment *String*	Segmentation *String*
	Ro'	CUSTOMERNAME	Recency	Frequency	Monetary	RFM_Segment	1 selected
2	Row1	Alpha Cognac	675	20	70,488.44	LLL	Lost Customer
11	Row...	Auto-Moto Classics Inc.	1353	8	26,479.26	LLL	Lost Customer
13	Row...	Bavarian Collectables Impor...	801	14	34,993.92	LLL	Lost Customer
21	Row...	Clover Collections, Co.	659	16	57,756.43	LLL	Lost Customer
30	Row...	Double Decker Gift Stores, L...	670	12	36,019.04	LLL	Lost Customer
36	Row...	Gift Ideas Corp.	947	19	57,294.42	LLL	Lost Customer
41	Row...	Iberia Gift Imports, Corp.	904	15	54,723.62	LLL	Lost Customer
50	Row...	Mini Auto Werke	717	15	52,263.9	LLL	Lost Customer
67	Row...	Royale Belge	737	8	33,440.1	LLL	Lost Customer
71	Row...	Signal Collectibles Ltd.	836	15	50,218.51	LLL	Lost Customer

CUSTOMER SEGMENTATION USING RFM ANALYSIS

Loyal Customers

These customers have consistently high Frequency and Monetary scores, reflecting their loyalty to the business.

#	RowID	CUSTOMERNAME *String*	Recency *Number (integer)*	Frequency *Number (integer)*	Monetary *Number (double)*	RFM_Segment *String*	Segmentation *String*
	Ro'	CUSTOMERNAME	Recency	Frequency	Monetary	RFM_Segment	1 selected
1	Row0	AV Stores, Co.	421	51	157,807.81	MHH	Loyal Customer
27	Row...	Danish Wholesale Imports	499	36	145,041.6	MHH	Loyal Customer
31	Row...	Dragon Souveniers, Ltd.	649	43	172,989.68	LHH	Loyal Customer
42	Row...	L'ordine Souveniers	493	39	142,601.33	MHH	Loyal Customer
56	Row...	Muscle Machine Inc	502	48	197,736.94	MHH	Loyal Customer
64	Row...	Reims Collectables	287	41	135,042.94	MHH	Loyal Customer
65	Row...	Rovelli Gifts	1032	48	137,955.72	LHH	Loyal Customer
69	Row...	Saveley & Henriot, Co.	586	41	142,874.25	MHH	Loyal Customer
70	Row...	Scandinavian Gift Ideas	262	38	134,259.33	MHH	Loyal Customer

CUSTOMER SEGMENTATION USING RFM ANALYSIS

Recommendations

- Best Customers: Offer exclusive perks, discounts, or loyalty rewards to retain these customers.
- Customers on the Verge of Churning: Re-engage through targeted email campaigns, offering personalized discounts or promotions.
- Lost Customers: Attempt to win them back with special “Miss You" campaigns or seek feedback to understand why they stopped engaging.
- Loyal Customers: Continue nurturing these relationships with VIP programs and regular communication to maintain their loyalty.

PART B

1. Problem Statement
2. Data Overview
3. Exploratory Data Analysis
    - Time – Series Analysis
    - Product Analysis
4. Market Basket Analysis
    - Relevance
    - KNIME Workflow Image
    - Threshold Values of Support & Confidence
5. Associations Identified
    - Metrics Explanation
6. Suggestion of Possible Combos & Offers
    - Recommendations
        - High-Frequency Item Combos
        - Complementary Product Discounts
        - Bundled Combos for Regular Items
        - Weekend Deals for Impulse Purchases
        - Buy One, Get One (BOGO) Offers
        - Seasonal Combos

PROBLEM STATEMENT

A grocery store shared the transactional data with you. Your job is to conduct a thorough analysis of Point of Sale (POS) data, identify the most commonly occurring sets of items in the customer orders, and provide recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.

DATA OVERVIEW

Dataset Contains one sheets namely [‘dataset group']. Which contains 3 Columns & 20641 Rows. There are no missing values. Date column need to altered to correct format.

EDA

Performed exploratory analysis and list out the insights from them.

Tool Used: KNIME

- Total No. Of Sales Records = 20641.
- Dataset have 3 Columns.
- No Missing Entries.
- Zero Duplicate Entries.
- Dataset has records from 1st Jan 2018 to 26th Feb 2020.
- There 37 Range of Products.
- Maximum transactions of 113 were made on 8th Feb 2019.
- Average number products brought in an order was 18.
- Store did not provided data for Q4 (Oct, Nov & Dec), So lets assume the store was closed during Q4.
- Poultry was highest ordered product.
- Hand soap was the least ordered product.

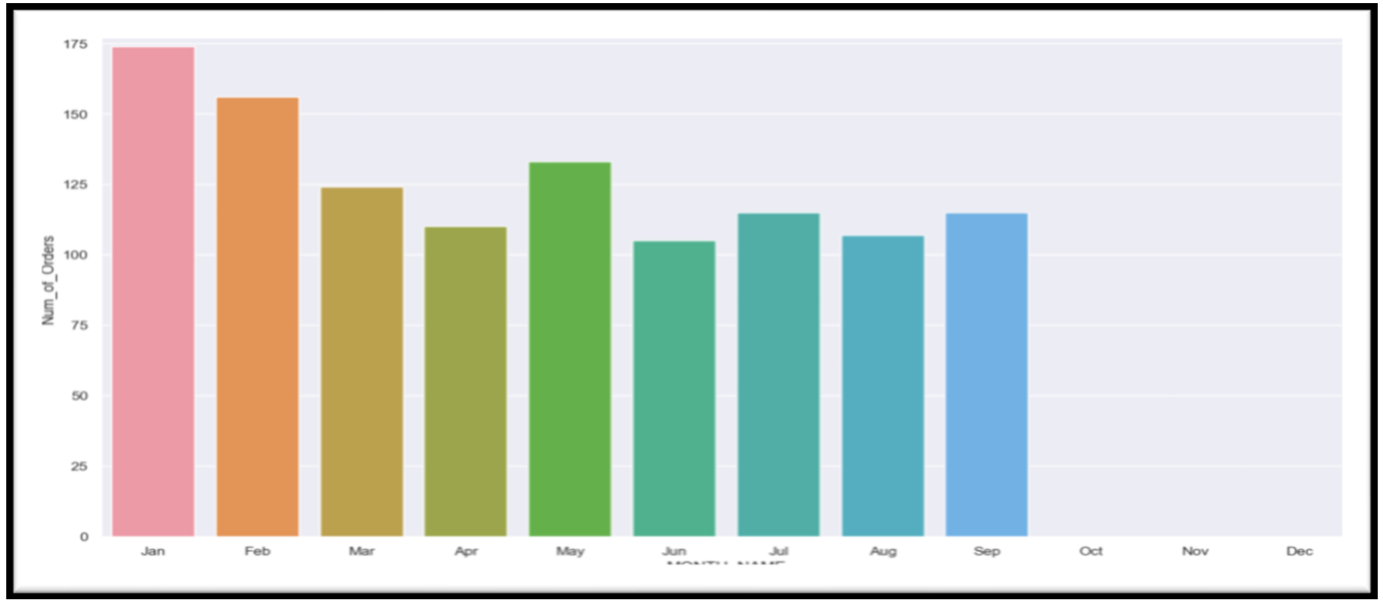
EDA

Statistical Summary:

- 1139 Number of Invoice and its retailed data was provided.

Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Mean
Date	Local Date	0	603	?	?	?	?	?	?
Order_id	Number (inte...	0	1139	1	1,139	292	581	862	575.986
Product	String	0	37	?	?	?	?	?	?

- Comparing of days of week Monday and Tuesday seems to have lowest transaction.
- Dataset do not have data 26th Feb 2020, may be store was closed because or Corona Pandemic.

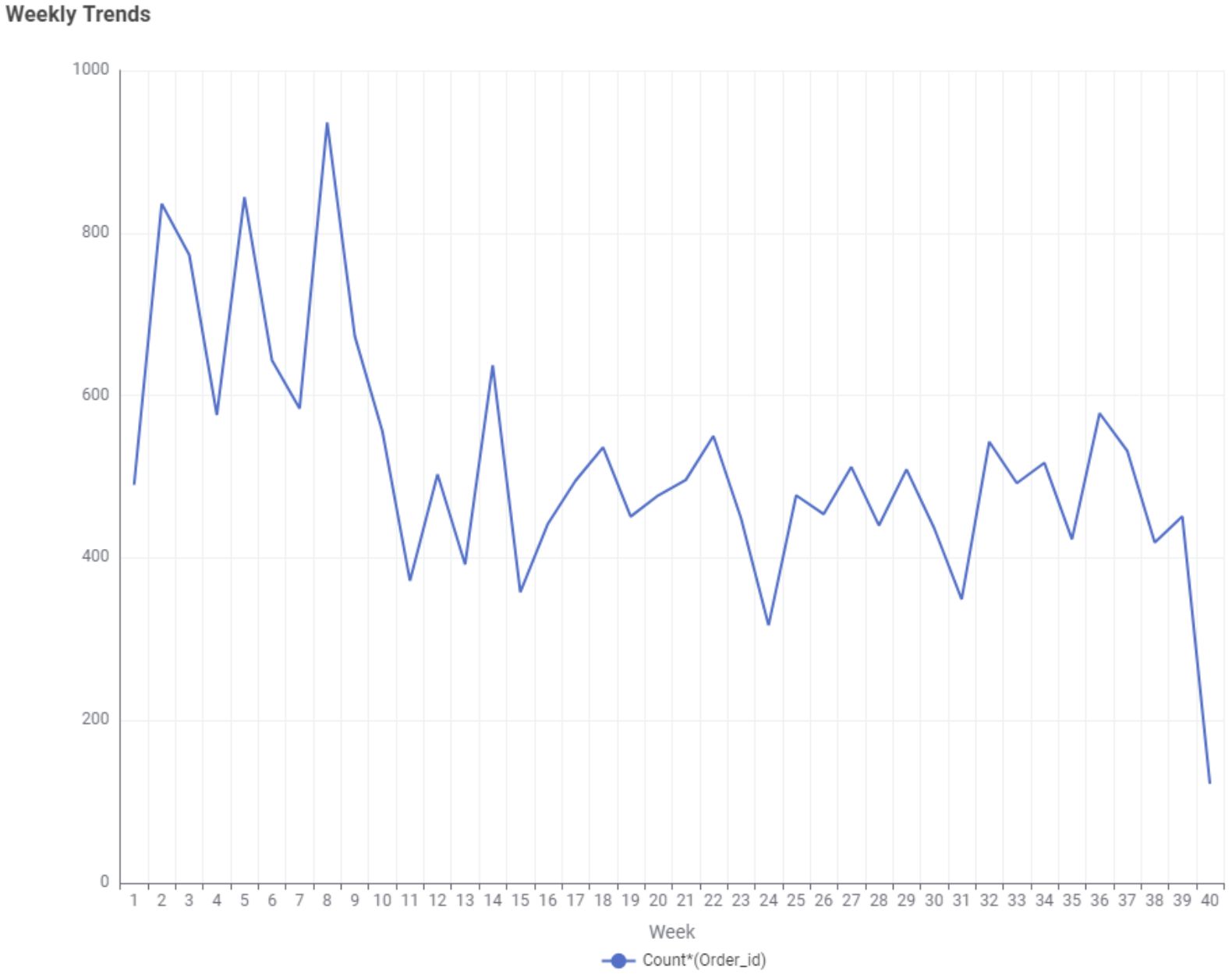


EDA

Insights from Weekly Trends:

• The metric experiences major week-to-week fluctuations, indicating high variability in the business performance, which could be influenced by factors like seasonality, promotions, or other external events.

• The overall trend shows a gradual decline in the metric over the time period shown.

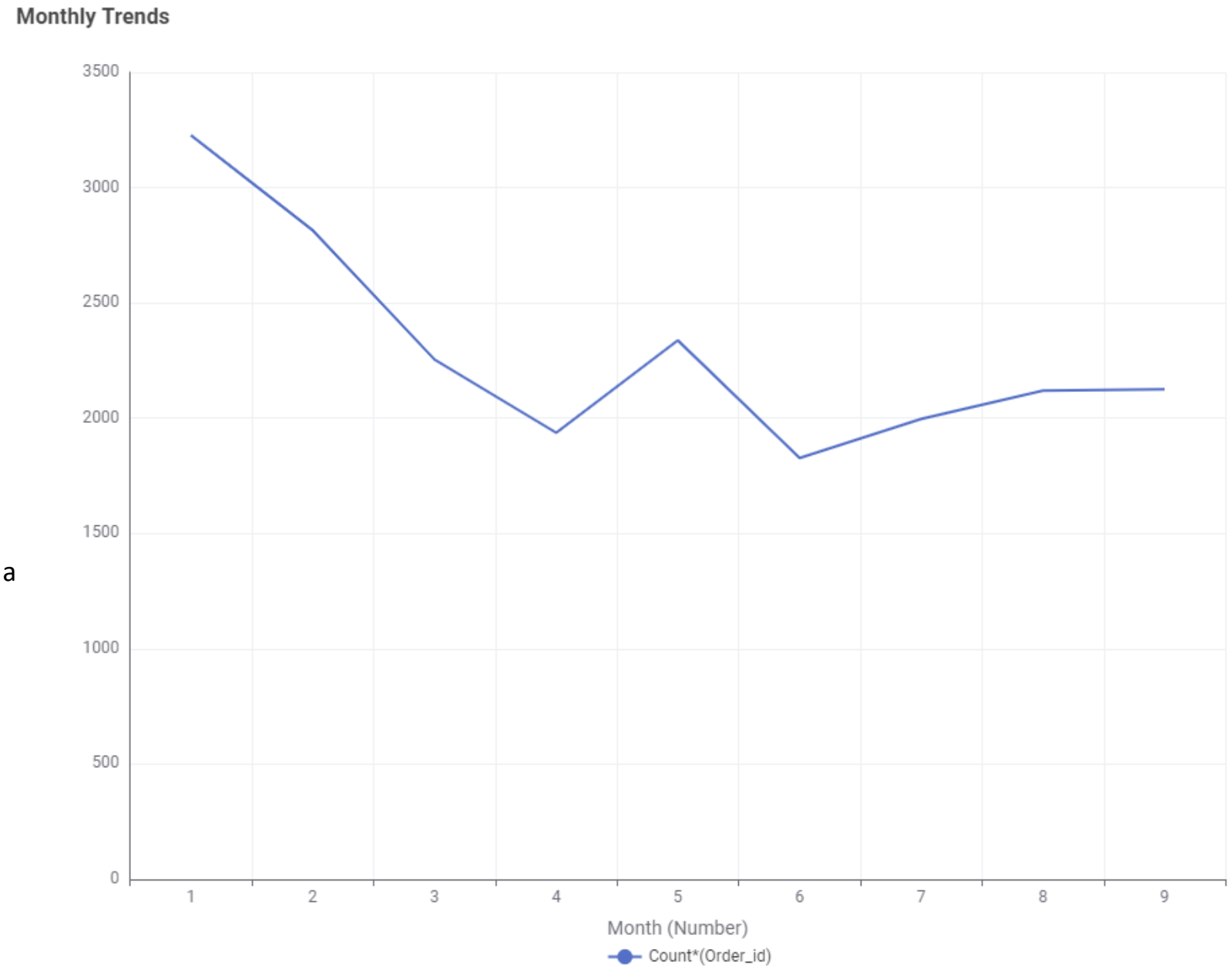


EDA

Insights from Monthly Trends:

• The metric exhibits large month-to-month variations, with sharp peaks and valleys throughout the time period.

• The overall trend shows a gradual decline in the metric's value, suggesting a downward trajectory in the business performance.



EDA

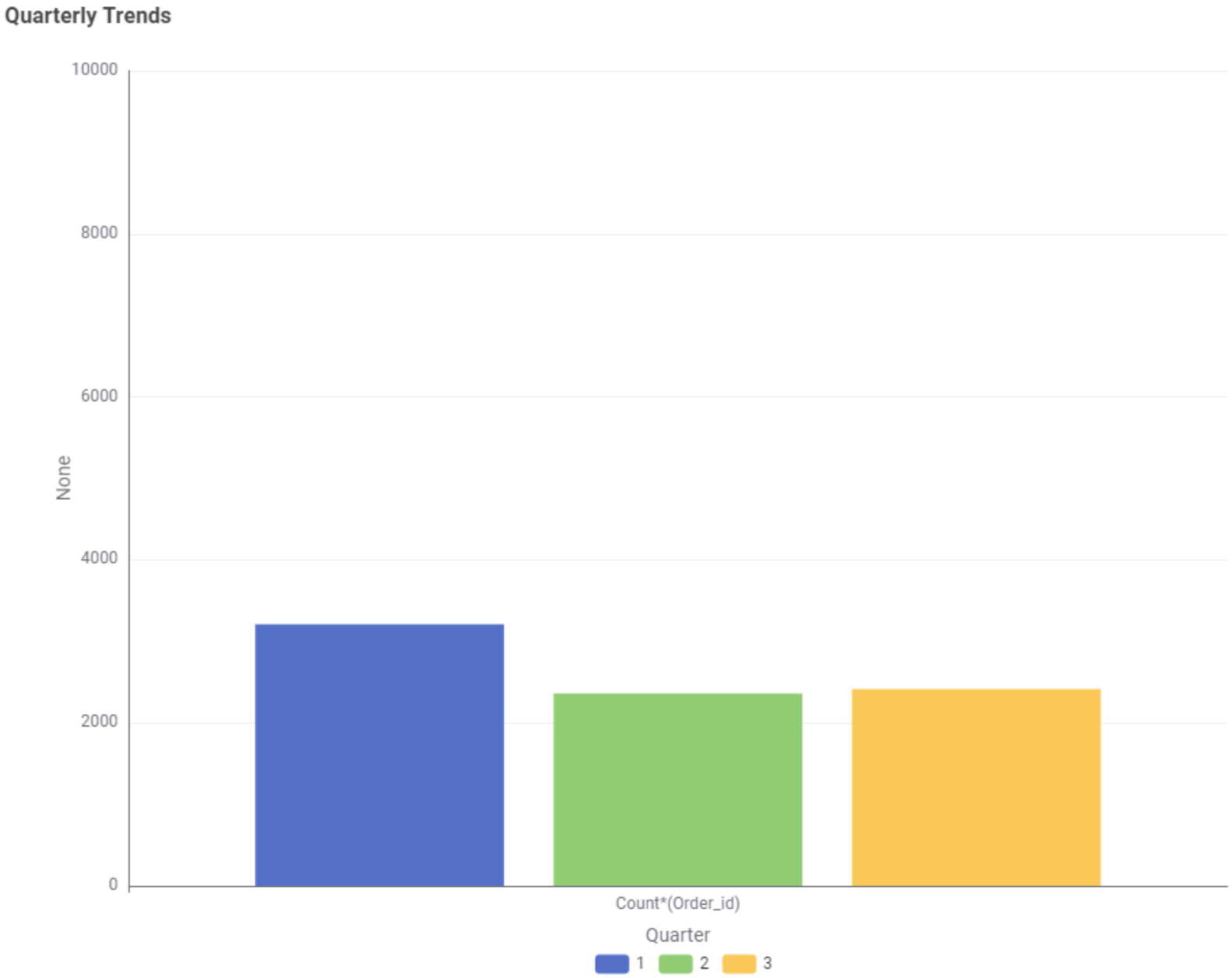
Insights from Quarterly Trends:

• The metric shows large differences in value across the three quarters presented, indicating substantial quarter-to-quarter fluctuations.

• The first quarter has the highest metric value, significantly higher than the other two quarters shown.

• The variations in quarterly values suggest the business may experience seasonal patterns, with a stronger performance in the first quarter compared to the others.

• Note – Q4 Data was not provided.



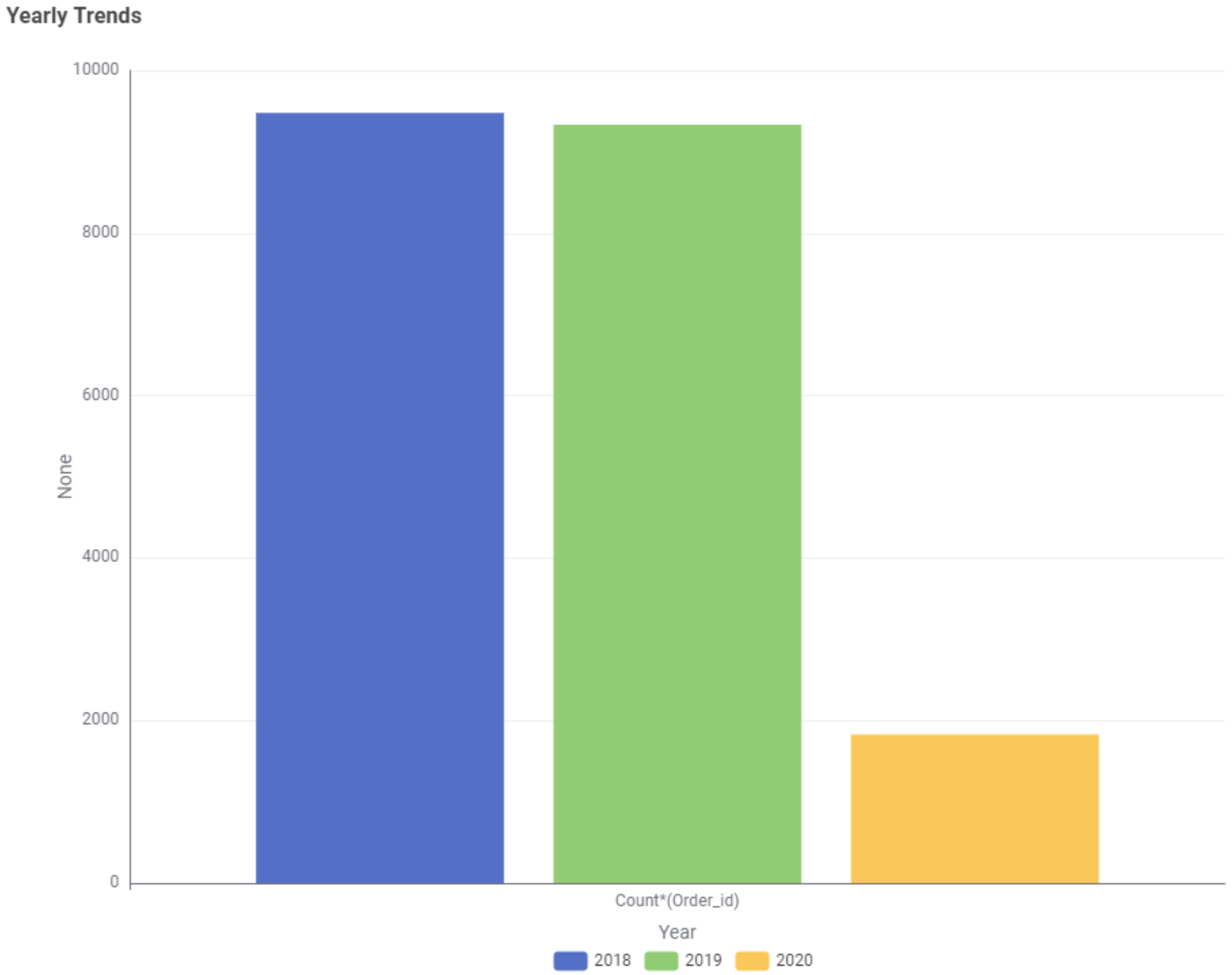
EDA

Insights from Yearly Trends:

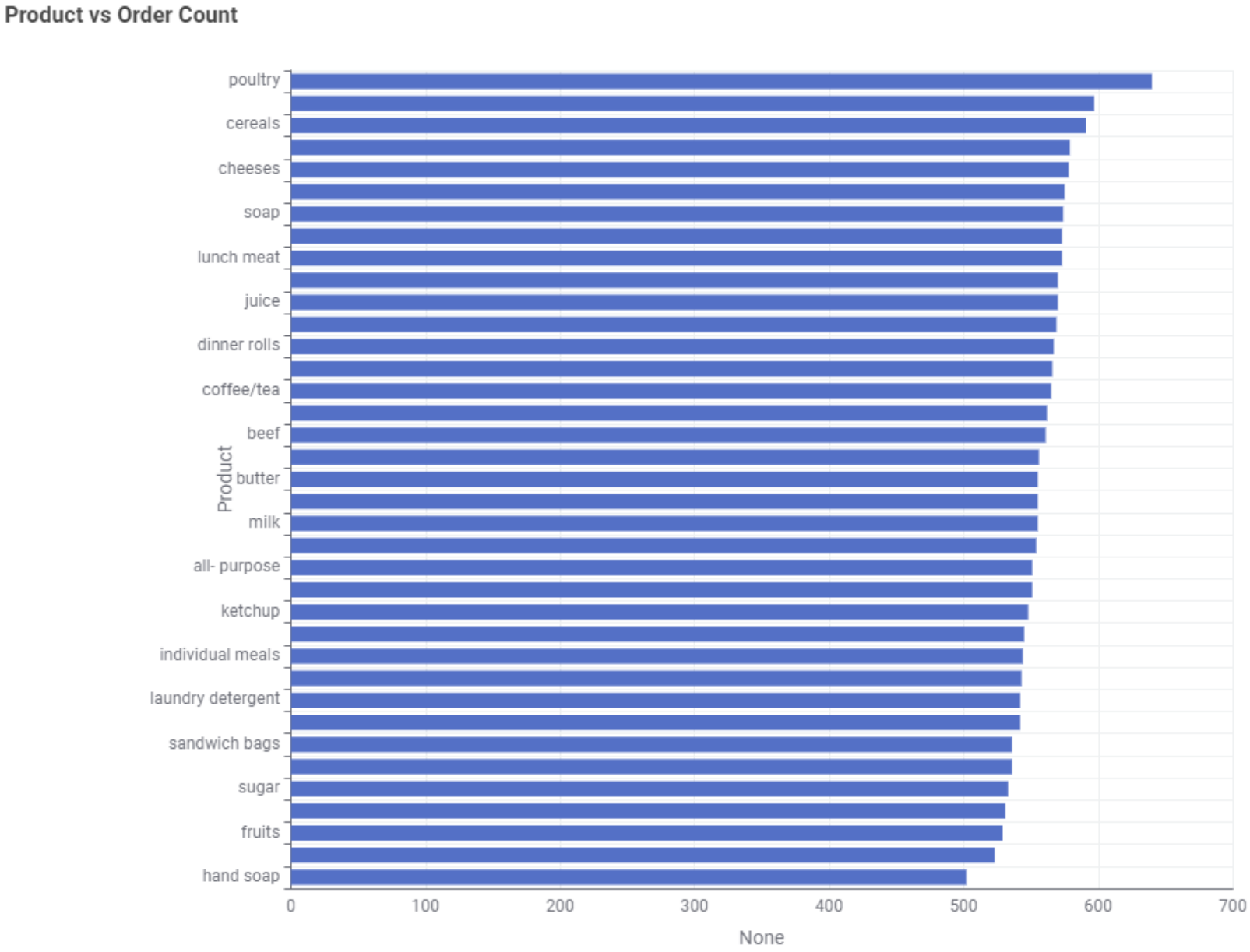
• The metric has Decreased substantially from 2018 to 2019, indicating Poor year-on-year performance improvement.

• Large Fall can be seen in 2020, that may be due to some external factor like Corona Pandemic.

• Note – Only first two-month data was available for year 2020.



EDA



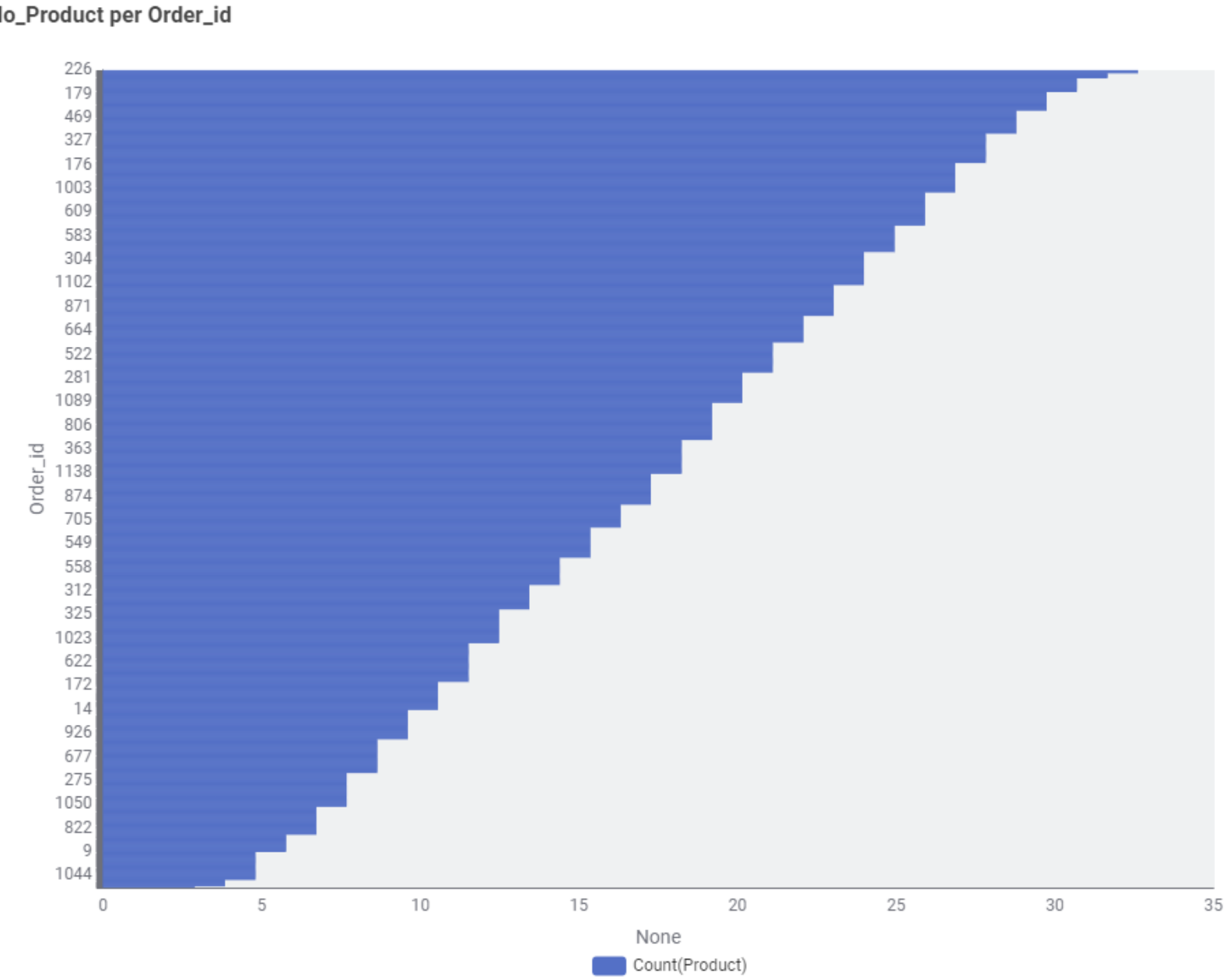
Key insights from the "Product vs Order Count" plot:

- The products with the highest order counts are poultry, cereals, and cheeses, indicating these are the top selling items.
- Categories like lunch meat, juice, and dinner rolls show moderate order counts, suggesting they are reasonably popular products.
- Items like sugar, fruits, and hand soap have relatively lower order counts, implying they may be slower moving products.

EDA

The plot shows the distribution of the number of products per order ID. Here are the key insights:

- The plot indicates that the majority of orders have a relatively high number of products, with the counts extending up to 34 products per order.
- The wide range of product counts per order implies there may be room to optimize the average basket size, either by encouraging customers to add more items or by identifying ways to reduce the number of single item orders.
- The high product counts per order could have implications for inventory management, warehouse operations, and logistics, as the business needs to ensure efficient order processing and fulfillment for these multi-item orders.



MARKET BASKET ANALYSIS

Market Basket Analysis (MBA) is a data mining technique used to uncover relationships or associations between items in a dataset. It helps identify products that are frequently bought together. In this case, MBA is crucial for understanding customer buying patterns in a grocery store.

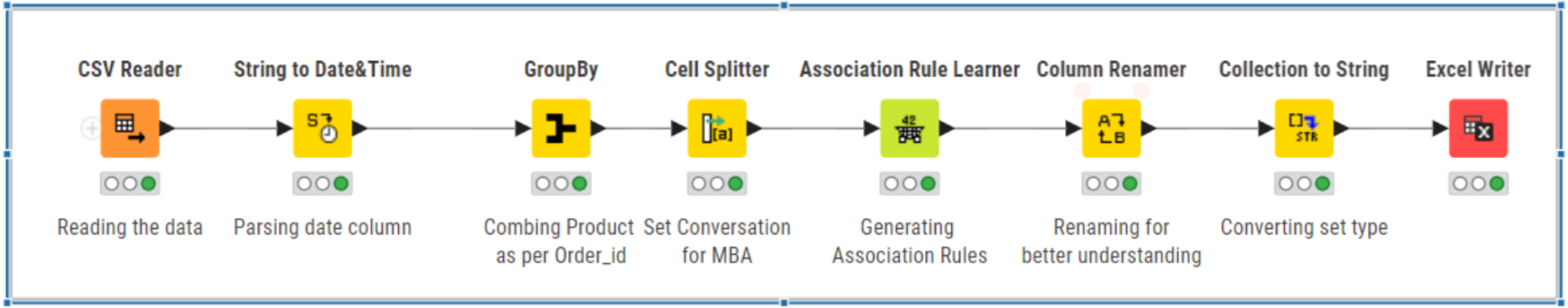
Relevance

• Combo Offers: By identifying frequently bought item sets, the grocery store can create combo offers or discounts that encourage customers to buy more.

• Store Layout: Items often purchased together can be strategically placed nearby to increase impulse buying.

• Personalized Promotions: Offers can be tailored based on customer purchase history to increase loyalty and sales.

KNIME Workflow Image



MARKET BASKET ANALYSIS

Threshold Values of Support & Confidence

Support:

- Represents how frequently an itemset occurs in the dataset.
- Example: A support of 0.05 means the itemset appears in 5% of transactions.
- Used Value: I used support value as 0.02 as it helps focus the analysis on the most frequent patterns in the data.

2. Confidence:

- Measures the likelihood of the second item being purchased when the first item is already in the basket.
- Example: A confidence of 0.8 means there is an 80% chance of buying item B given item A was purchased.
- Used Value: I used 0.65 as it helps focus the analysis on the more reliable and actionable association rules.

ASSOCIATIONS IDENTIFIED

The associations identified, below is output from the “Association Rule Learner Node” output:

#	RowID	Support *Number (double)*	Confidence *Number (double)*	Lift *Number (double)*	Consequent *String*	implies *String*	Items *Set*
1	rule0	0.02	0.657	1.786	sandwich bags	<---	[yogurt,mixes,pork,...]
2	rule1	0.02	0.676	1.731	soda	<---	[yogurt,waffles,pork,...]
3	rule2	0.02	0.657	1.729	milk	<---	[yogurt,poultry,pork,...]
4	rule3	0.02	0.697	1.877	ketchup	<---	[yogurt,pork,juice,...]
5	rule4	0.02	0.793	2.151	shampoo	<---	[yogurt,dinner rolls,lunch m...
6	rule5	0.02	0.676	1.739	dinner rolls	<---	[yogurt,all- purpose,flour,...]
7	rule6	0.02	0.697	1.859	all- purpose	<---	[yogurt,butter,laundry deter...
8	rule7	0.02	0.657	1.737	laundry detergent	<---	[yogurt,butter,all- purpose,...]
9	rule8	0.02	0.697	1.654	poultry	<---	[yogurt,all- purpose,tortillas,...
10	rule9	0.02	0.657	1.745	juice	<---	[yogurt,beef,pork,...]
11	rule10	0.02	0.697	1.784	cheeses	<---	[yogurt,dinner rolls,coffee/t...
12	rule11	0.02	0.657	1.745	juice	<---	[yogurt,shampoo,pork,...]
13	rule12	0.02	0.657	1.749	mixes	<---	[yogurt,poultry,pork,...]
14	rule13	0.02	0.657	1.66	cereals	<---	[yogurt,coffee/tea,pork,...]
15	rule14	0.02	0.657	1.733	coffee/tea	<---	[yogurt,cereals,pork,...]
16	rule15	0.02	0.676	1.716	waffles	<---	[yogurt,toilet paper,pork,...]
17	rule16	0.02	0.676	1.796	juice	<---	[yogurt,waffles,toilet paper,....
18	rule17	0.02	0.676	1.731	cheeses	<---	[yogurt,toilet paper,coffee/t...
19	rule18	0.02	0.719	1.895	coffee/tea	<---	[yogurt,cheeses,toilet paper...
20	rule19	0.02	0.697	1.838	coffee/tea	<---	[yogurt,milk,tortillas,...]

ASSOCIATIONS IDENTIFIED

Explain about support, confidence, & lift values that are calculated

Support:

- Measures how frequently the itemset (antecedent and consequent) appears in the dataset.
- Formula: Support=Transactions with Antecedent and Consequent/(TotaL Transaction)
- Interpretation: A support of 0.02 for "Yogurt → Sandwich Bags" means 2% of all transactions include both items.

Confidence:

- Represents the probability of purchasing the consequent item(s) given that the antecedent item(s) are purchased.
- Formula: Confidence=Transactions with Antecedent and Consequent/(Transaction with Antecedent)
- Interpretation: A confidence of 0.65 for "Yogurt → Sandwich Bags" indicates that 65% of customers who bought yogurt also bought sandwich bags.

Lift:

- Measures the strength of the rule over random co-occurrence of items.
- Formula: Lift=Confidence/(Support of Consequent)
- Interpretation: A lift greater than 1 implies a positive association between the items. For "Yogurt → Sandwich Bags," a lift of 1.78 indicates that the items are 78% more likely to be bought together than by chance.

SUGGESTION OF POSSIBLE COMBOS & OFFERS

Recommendation

Based on the associations identified in the Market Basket Analysis, here are specific combo suggestions and promotions that the grocery store can implement to drive sales and improve customer satisfaction:

1. **High-Frequency Item Combos**

Item combinations with the highest Support. Below are the top 10 list,

Support	Confidence	Lift	Item Recommendation	implies	Item Basket
0.053555751	0.655913978	1.556429211	poultry	<---	[dinner rolls,spaghetti sauce,laundry detergent]
0.051799824	0.686046512	1.627931202	poultry	<---	[dinner rolls,spaghetti sauce,ice cream]
0.050921861	0.674418605	1.726208518	cheeses	<---	[bagels,cereals,sandwich bags]
0.049165935	0.651162791	1.633644094	ice cream	<---	[paper towels,eggs,lunch meat]
0.049165935	0.674698795	1.601004016	poultry	<---	[beef,toilet paper,sugar]
0.049165935	0.658823529	1.563333333	poultry	<---	[dinner rolls,spaghetti sauce,sugar]
0.048287972	0.6875	1.631380208	poultry	<---	[dinner rolls,spaghetti sauce,soap]
0.048287972	0.654761905	1.683462324	dinner rolls	<---	[spaghetti sauce,poultry,hand soap]
0.048287972	0.662650602	1.572414659	poultry	<---	[dinner rolls,spaghetti sauce,hand soap]

2. **Complementary Product Discounts**

Item pairs with high Confidence and moderate Lift. These indicate strong but not overly coincidental relationships. Below was top 10 list that has high confidence and moderate lift (lift >1.5 &<2).

Support	Confidence	Lift	Item Recommendation	implies	Item Basket
0.02809482	0.842105263	1.998245614	poultry	<---	[dinner rolls,spaghetti sauce,beef,sugar]
0.022827041	0.838709677	1.990188172	poultry	<---	[dinner rolls,spaghetti sauce,hand soap,soap]
0.025460931	0.828571429	1.966130952	poultry	<---	[shampoo,hand soap,juice,sugar]
0.021071115	0.827586207	1.963793103	poultry	<---	[eggs,tortillas,coffee/tea,sugar]
0.02809482	0.820512821	1.947008547	poultry	<---	[dinner rolls,spaghetti sauce,sandwich loaves,soap]
0.022827041	0.8125	1.927994792	poultry	<---	[spaghetti sauce,laundry detergent,mixes,sugar]
0.021949078	0.806451613	1.913642473	poultry	<---	[butter,cheeses,sandwich loaves,laundry detergent]
0.025460931	0.805555556	1.911516204	poultry	<---	[dinner rolls,spaghetti sauce,ice cream,beef]
0.024582968	0.8	1.898333333	poultry	<---	[dinner rolls,spaghetti sauce,sandwich loaves,hand soap]
0.024582968	0.8	1.898333333	poultry	<---	[dinner rolls,spaghetti sauce,hand soap,coffee/tea]
0.02809482	0.8	1.898333333	poultry	<---	[dinner rolls,spaghetti sauce,laundry detergent,sugar]

SUGGESTION OF POSSIBLE COMBOS & OFFERS

3. **Bundled Combos for Regular Items**

Target pairs with both high Support and Lift, as these are strong associations. Below was list of items that has support > 0.05 and lift >1.5.

Support	Confidence	Lift	Item Recommendation	implies	Item Basket
0.050921861	0.674418605	1.726208518	cheeses	<---	[bagels,cereals,sandwich bags]
0.051799824	0.686046512	1.627931202	poultry	<---	[dinner rolls,spaghetti sauce,ice cream]
0.053555751	0.655913978	1.556429211	poultry	<---	[dinner rolls,spaghetti sauce,laundry detergent]

4. **Weekend Deals for Impulse Purchases**

Item pairs with high Confidence and moderate Lift. These indicate strong but not overly coincidental relationships. Below was top 10 list that has high confidence and moderate lift (lift >1.5 &<2).

Support	Confidence	Lift	Item Recommendation	implies	Item Basket
0.020193152	0.884615385	2.21933243	ice cream	<---	[paper towels,eggs,dinner rolls,pasta,lunch meat]
0.026338894	0.857142857	2.193900482	cheeses	<---	[paper towels,cereals,sandwich bags,sugar]
0.020193152	0.851851852	2.349296027	paper towels	<---	[eggs,dinner rolls,ice cream,pasta,lunch meat]
0.020193152	0.851851852	2.266960886	mixes	<---	[yogurt,dishwashing liquid/detergent,all- purpose,hand soap]
0.020193152	0.851851852	2.180357886	soda	<---	[bagels,pasta,individual meals,pork]
0.028972783	0.846153846	2.007852564	poultry	<---	[dinner rolls,spaghetti sauce,hand soap,sugar]
0.02809482	0.842105263	1.998245614	poultry	<---	[dinner rolls,spaghetti sauce,beef,sugar]
0.022827041	0.838709677	2.258369557	ketchup	<---	[tortillas,coffee/tea,juice,soap]
0.022827041	0.838709677	1.990188172	poultry	<---	[dinner rolls,spaghetti sauce,hand soap,soap]
0.021949078	0.833333333	2.243892829	pasta	<---	[paper towels,dishwashing liquid/detergent,eggs,dinner rolls,ice cream]

SUGGESTION OF POSSIBLE COMBOS & OFFERS

5. **Buy One, Get One (BOGO) Offers**

Combinations where one product strongly implies another (Confidence > 0.8) and has a high Lift value, below are list of those combinations (Confidence > 0.8 and Lift > 1.5)

Support	Confidence	Lift	Item Recommendation	implies	Item Basket
0.020193152	0.851851852	2.349296027	paper towels	<---	[eggs,dinner rolls,ice cream,pasta,lunch meat]
0.020193152	0.851851852	2.266960886	mixes	<---	[yogurt,dishwashing liquid/detergent,all- purpose,hand soap]
0.020193152	0.821428571	2.265392598	paper towels	<---	[eggs,dinner rolls,poultry,ice cream,pasta]
0.022827041	0.838709677	2.258369557	ketchup	<---	[tortillas,coffee/tea,juice,soap]
0.021949078	0.833333333	2.243892829	pasta	<---	[paper towels,dishwashing liquid/detergent,eggs,dinner rolls,ice cream]
0.020193152	0.884615385	2.21933243	ice cream	<---	[paper towels,eggs,dinner rolls,pasta,lunch meat]
0.021071115	0.827586207	2.217931034	spaghetti sauce	<---	[waffles,laundry detergent,mixes,soap]
0.021071115	0.827586207	2.207542599	beef	<---	[poultry,fruits,hand soap,sugar]
0.026338894	0.857142857	2.193900482	cheeses	<---	[paper towels,cereals,sandwich bags,sugar]
0.020193152	0.821428571	2.191117431	beef	<---	[shampoo,fruits,lunch meat,pork]

6. **Seasonal Combos**

Item pairs where Support is moderate but the combination is unique (high Lift).

Support	Confidence	Lift	Item Recommendation	implies	Item Basket
0.030728709	0.795454545	2.19375963	paper towels	<---	[eggs,ice cream,pasta,lunch meat]
0.031606673	0.782608696	2.158332456	paper towels	<---	[eggs,ice cream,pasta,cereals]
0.030728709	0.729166667	2.065972222	flour	<---	[dishwashing liquid/detergent,cheeses,waffles,soda]
0.032484636	0.74	2.040823245	paper towels	<---	[eggs,dinner rolls,ice cream,pasta]
0.031606673	0.72	1.98566586	paper towels	<---	[eggs,poultry,ice cream,pasta]
0.030728709	0.777777778	1.951297112	ice cream	<---	[paper towels,eggs,pasta,lunch meat]
0.030728709	0.760869565	1.947484123	soda	<---	[dishwashing liquid/detergent,cheeses,flour,waffles]
0.033362599	0.716981132	1.930594585	pasta	<---	[paper towels,dishwashing liquid/detergent,eggs,ice cream]
0.040386304	0.696969697	1.922151295	paper towels	<---	[all- purpose,individual meals,toilet paper]
0.030728709	0.714285714	1.914285714	spaghetti sauce	<---	[dinner rolls,poultry,laundry detergent,juice]

THANK YOU